The Role of the HCD/CAIAT Project in Improving the Ability of Science Teachers for Constructing HCD Test Items in the Kingdom of Saudi Arabia.

By

Mohammed Ibraheem A. Al-Mulhim

Submitted in fulfilment of the requirement for the Degree of Doctor of Philosophy (DPhil) in Education.

School of Education

University of Sussex

Feb 2012

I hereby declare that this thesis has not been, and will not be, submitted in whole or in part to another University for the award of any other degree and that it has not been previously submitted to this University for a degree.

Signature : .....

Dedication

This work is dedicated to:

my mother Muneerah for her prayers for me,
my father Ibraheem (May Allah bless him) for his prayers for me,
my wife Norah for her support, her patience and her appreciation,
and my kind children… for their true-hearted wishes for me.

May Allah bless them all.

Acknowledgments

I would like to thank the University of Sussex for giving me this opportunity to complete my PhD at such a famous and respected institution. Specifically, my thanks to the university for the good service of the electronic library that enabled me as a student working abroad to gain access to the most important and rich sources of information. I would also like to thank the London Institute of Education, Kings College, King Saud University, Um Al-Qura University, King Abdulaziz University and King Fahad National Library for the good service that I have received during my personal visits to their libraries.

I would like to express a big thank you to my supervisor, Professor Keith Lewin, who showed his care, sincere advice and support. When my job and/or social circumstances caused some delay, he was always sympathetic and supportive. I worked my way through many crisis situations in terms of exceeding registration time limits or being in need of an extension, and throughout all of these I found Keith provided all the support possible. Above all, his comments, critiques, indications, explanations and suggestions were rich and informative to the extent of transforming my ability in each successive stage of the work to a higher level of research quality performance. His insightful feedback on my work was invaluable. Also, my great thanks to Professor Judy Sebba from Sussex University for her orientation and advice during the drawing up of my research plan, Philip Adey from Kings College for his positive response to my queries about CASE intervention, as well as Peter Adamczyk from Sussex University for his sincere response and dedicated assistance when I visited him during my study about CASE intervention.

My grateful thank to the Ministry of Education in KSA for its support for developmental projects such as the present one, which encouraged me to undertake this research;, the General Manager of Boys' Education at Al-Ahsa; Ahmed Al Ghunaim, for his support during the pilot stage; the Physics Educational Supervisors who worked with me in the pilot stage of the project as research assistants, as well as the female Science Educational Supervisors from

the General Directorate for Girls' Education at Al-Ahsa for taking the role of research assistants at the second application of the field work with the main sample. My thanks, too, to Khalid Al-Mansour for his valuable contribution in arranging and following up the training course requirements, to Hesham Al-Thunayan and Ahmed AL-Musallam for their cooperation in setting up the training lab for training, to Hani Al-Shayeb for his technical support for the participating teachers at all stages of the field work, to Iqbal Al Shuhail for translating the film about CASE intervention into Arabic, printing the translated script into the film scenes and translating the form for performance evaluation criteria from Arabic to English, to Al-Khaldeyyah Private Schools for sponsoring the training activities of the field work, to my brother Khalid Al-Mulhim for his contribution in initial proof-reading of this report, to my sister Hanan Al-Mulhim for designing a logo for CAIAT software, and to my daughters Muneerah and Yasmeen for their help in making some data entries for the research report.

My special thanks go to the Head of the Measurements and Evaluation Unit at the General Directorate for Girls' Education at Al-Ahsa, Kholood Al-Kulaibi. I would like to record my appreciation of her enormous contribution to the fieldwork. She worked hard in leading the team of research assistants to get the fieldwork with female teachers off the ground, and remained at the hub of the process, contacting me, asking, learning and reporting until the end. She also contributed to reviewing the project's user manuals, suggesting some important additions, and supervised the process of entering raw data into SPSS.

I would like to show my true-hearted and unending thanks to my wife, Norah Al Mudhaffer, for being an endlessly helpful and caring human being at the centre of my hard work on this research for more than seven years. I owe so much to Norah's patience, prayers, appreciation and passionate wishes for my success. Also, my children who shared Norah's patience in their response to my busy time of research work alongside my heavy workload as a general manager. They suffered from not travelling and not receiving enough support from me for their study. They all remain in my heart forever.

Finally, my heartfelt appreciation to my parents for their unflagging love and support throughout my life; this work would have been simply impossible without them. I am indebted to my mother, Muneerah, for her continual prayers, wise advice, and educational way of rearing me despite not having had the advantages of a formal education herself. Her struggle in this life for her children's success and her family's prosperity is spectacular. Although it is common in my society for people to feel proud of their fathers, I feel proud of my mother the same as I feel proud of my great father, a man of true faith and great kindness. My mother's support and everlasting love will be the moon that lights up my way for the rest of my life. I am grateful to my father, Ibraheem, who, as a typical father in a Saudi family, worked industriously to provide the best possible environment for his children to grow up in and attend school. He provided me with a great example of the believer who looks at this life as simply as possible, with a wisdom that many lack. I have gained invaluably by following his vision, belief and sincere worship, which exceeded all that I learnt in school and to which I attribute whatever success I have had. Although he is no longer with us, he is forever remembered, and I am sure that in the heaven of his grave, he shares our joy and happiness.

Table of Contents

List of Acronyms and Abbreviations

ALESCO: Arab League Educational, Scientific and Cultural Organisation

APA: American Psychological Association

AR: Action Research

ASC: Assessment Systems Corporation

CA: Cognitive Acceleration

CAIAT: Computer Aided Item Analysis Technique

CASE: Cognitive Acceleration through Science Education

CPD: continuing professional development

CS: Case Study

CTT: Classical Test Theory

D: Discrimination Index

DEME: Department for Educational Measurement and Evaluation

EMF: Educational Measurement Framework

ERIC: Educational Resources Information Centre

GCSE: General Certificate for Secondary Education

GDBEA: General Directorate of Boys Education at Al-Ahsa

GDEME: General Directorate of Educational Measurements and Evaluation

GDER: General Directorate of Educational Research

GDGEA: General Directorate of Girls Education at Al-Ahsa

HCD: Higher Cognitive Demand

IAT: Item Analysis Technique

ICC: Item Characteristic Curve

ICT: Information and Computer Technology

INSET: In Service Training

IRT: Item Response Theory

KSA: Kingdome of Saudi Arabia

LCD: Low Cognitive Demand

Lertap: Laboratory of Educational Research Test Analysis Package

LEA: Local Education Authority

MOE: Ministry of Education

NC: National Curriculum

PC: Personal Computer

P: Difficulty Coefficient

PD: Professional Development

PF: Psychometric Framework

PRESET: Pre Service Training

SATs: Standard Assessment Tasks (UK)

SAT: Scholastic Aptitude Test and Scholastic Assessment Test (USA)

SBA: School-Based Assessment

SEAC: Secondary Examinations and Assessment Council

SEN : Special Educational Needs

TA: Teacher Assessment

TGAT: Task Group on Assessment and Testing

UK: United Kingdome

UNESCO: United Nations Educational, Scientific and Cultural Organisation

USA: United States of America

ZPD: Zone of Proximal Development

List of Tables

List of Figures

List of Diagrams

Abstract

The objective of this research is to participate in improving the quality of education in the Kingdom of Saudi Arabia (KSA) by developing the skills of Saudi female science teachers in writing higher cognitive demand (HCD) questions of exemplary quality. It is an evaluative study that follows the descriptive method of research design by depending on a combination of both quantitative and qualitative inquiry. Therefore, various instruments for collecting data were employed. The sample size of 409 represents all of the female science teachers who work in the girls' schools in the urban area of Al Ahsa, a city in KSA.

A suggested program called HCD/CAIAT is introduced and the main objective of the present evaluative research is to examine this project's functional potential to improve the researched sample related practices. The project includes an innovative software package, the Computer Aided Item Analysis Technique (CAIAT) designed purposely for this research in the Arabic language to provide the sample teachers with the two parameters of classical item analysis that indicate the strengths or weaknesses of a test question (difficulty and discrimination). This package is introduced through a training course that also trains the teachers in skills of question construction and teaching on HCD level. The CAIAT is intended to stimulate the teachers' professional development (PD) by raising their awareness of the validity of their HCD test items and encouraging/assisting them to improve their HCD questions over time which is anticipated to help improve their instruction. This concept of utilising CAIAT for improving teachers' practices is breaking new ground and establishing a basis for further development in the field of study.

The main purpose of the research is to answer the following two major questions. The first is to what extent can the HCD/CAIAT project assist female science teachers in Saudi schools to improve their ability to analyse their test questions, so as to write exemplary HCD test items and to teach at HCD level (*Effectiveness* dimension)? And the second is, to what extent could this be reflected in their on-going practice both for the test construction and for teaching (*Adoption* dimension)?

The findings have indicated that the sample teachers' prior background in the researched concepts and skills (HCD and IAT) are limited. However, the effectiveness dimension findings showed that the teachers have successfully acquired all of the

project's abilities/skills: knowledge of HCD concepts, skills of writing HCD instructional objectives and HCD questions, and using/utilising CAIAT successfully for assessing their test items. For the adoption dimension, the HCD/CAIAT package was successful in encouraging the teachers to adopt HCD and IAT which was a result of the successful role of the CAIAT software in stimulating the teachers' PD for learning (on their own) how to improve their assessment skills for HCD levels.

Furthermore, the research has identified ten study variables, which are the teachers' background characteristics, in order to test the statistical significance of their role in the reported differences amongst the results found for the various aspects measured by the research data collection instruments. These teachers' characteristics are: educational qualification, prior training on test construction skills, prior training on IAT, key stage (intermediate/secondary), level of graduation (GPA/equivalent), years of experience in teaching, specialisation subject, prior experience in using computers, possess of a PC at home and ability to use some mainstream software packages. Statistically, the impact of these variables on the teachers' acquisition or adoption of the project's concept and skills was found very limited; which supports generalizability of the research findings. It is recommended that the Ministry of Education (MoE) at KSA adopt the HCD/CAIAT package in order to encourage all KSA female science teachers to tackle HCD levels in their instruction and assessment, which is very likely to have a positive impact on their efforts in teaching thinking and inducing creativity. Ten other recommendations were also suggested.

Keywords: Kingdom of Saudi Arabia (KSA), Higher Cognitive Demand (HCD), Item Analysis Technique (IAT), Computer Aided Item Analysis Technique (CAIAT), Professional Development (PD)

Chapter 1

Introduction[1]

> "Teachers should remember that while there is no such thing as a perfect test, we should strive continually to achieve this ideal" (Collins et al., 1976, p. 107)

1.1 Preface

Educational assessment has become one of the most important areas of educational reform in recent years. Many reform and development activities arise from assessment and imply that it is a major component: "Many educators and policymakers believe that what gets assessed is what gets taught and that the format of assessment influences the format of instruction" (Bond, 1994). Boersma (1976) found that teachers who were systematically involved in the construction of tests had a clearer perception of the curriculum; similarly, Bloom (1961) noted that these particular teachers conveyed more relevant instruction to their pupils (Stanley and Hopkins, 1978, p. 173). Williams (1991) presented research indicating that teachers spend approximately 15% of their time testing, and between 25% and 33% of their time measuring student achievement through classroom testing; and that they write 65.6% of their own test questions, obtaining the remainder from test guides and the like. Teachers view tests as important instructional tools that are worth the time and effort required for their implementation. Furthermore, in 1984, Benjamin Bloom found in his semi-experimental research of teaching pedagogy, and through a number of suggested modules, that the extensive use of classroom assessment was a key practice for better learning (Stiggins, 2002, p. 7).

This research taps into the issue of teachers assessing their pupils as part of a formative assessment, and considers this assessment vital to the quality of the teaching profession – and hence the quality of resulting educational outcomes. This study appreciates notions of educational change and modern approaches to learning, and it is formulated in a project style in order to introduce a newly innovated instrument (computer software), which is supported by training and administrative follow-up, to support a new concept for teachers' professional development that is anticipated to contribute to increasing the quality of education. By using an evaluative research approach, this study will seek to examine the extent to which such intervention could lead to better teaching practice in the fields of assessment and pedagogy.

In this chapter, I refer mainly to local literature to show the extent to which this research problem is significant in the light of previous related works. However, in the chapters that follow, especially those introducing the theoretical background; that is, Chapters 2, 3 and 4, I have referred to Western literature, especially that originating from the UK and USA. This is due to several reasons. First, there is a lack of rich and well-structured literature in the Saudi context or in the Arabian region. Although a number of Saudi research reports on education have been written, for postgraduate degrees or by university academic staff members, compared to Western sources and publications these are less diverse and therefore do not provide all the possible forms of knowledge to inform the sort of multidimensional research undertaken here. The majority of these Saudi or Arabian papers are at Masters degree level and are thus very likely to lack pure originality, although they are valid for their purpose. Moreover, the few studies that are relevant mostly use conceptual structures that are drawn from the Western/global literature. In fact, there is no established alternate analytical tradition in the existing literature that could be used as a conceptual base.

Second, the themes in Western literature to which I shall refer are of a general nature, such as learning, change or assessment, concepts which I believe are generally similar across the globe. Lessons learnt from such general issues in the West or the Far East could be utilised readily for other developing countries in the Middle East such as the Kingdom of Saudi Arabia (KSA). Third, in the past, the KSA's educational culture has benefited mainly from the British and the American educational experiences, which are represented clearly in areas of curriculum, assessment and development[2]. This in turn facilitates and justifies the intensive use of Western literature in framing new approaches for improving education in the KSA.

1.2 Overview of the Report Structure

In the present chapter, I will introduce the research problem by defining its predominant terminologies, statement of the problem, and aim of the study. As a background to the rationale of the study, I will present a review of the literature on the three themes underpinning the research concept. This review covers general and local literature, and the latter shows the extent to which the problems tackled by this research are reported in KSA. An elaboration on the rationale of the study will follow, then previous similar or relevant studies will be illustrated to provide an insight into the role

of this research in enriching this field and how, conceptually, it could be integrated with counterpart efforts. After this, I will explain the significance of this research, showing the educational/academic merit, originality, and long and short-term benefits to various stakeholders. Finally, and for the benefit of readers who have only a limited knowledge of Saudi education, I will provide, with some elaboration, key contextual information on KSA and its education.

In Chapter 2, *Learning and Science Education*, major approaches applied throughout the history of educational psychology will be illustrated by highlighting a number of well-known theories underpinned by these approaches. I will present styles of learning and teaching in KSA to provide an insight about where the Saudi school's common practices are located in this respect. Then, I will elaborate on learning in science education followed by presenting a number of studies that include related attempts at change within science education and highlight the lessons that could be learnt from these attempts.

Chapter 3, *Educational Change*, will give an overview about the nature of change in general and its role in education in particular. Perspectives of change will be demonstrated, focusing on organisational change, while personal change will also be highlighted because the individual's role in change is a substantial one. In this respect, the emotional dimension of personal change and two motivational theories will be explained, and professional development (PD) as the paradigm for the change envisaged by this project will be defined. In particular, learning with reflection and action research (AR) are two important themes of PD that will be given further elaboration. Since this research includes an innovative product (the CAIAT computer software), the *Diffusion of Innovation* (DoI) theory is demonstrated as a framework that gives further insight into the adoption of innovation. Finally, change in KSA is explained in the light of three national studies and connected to the DoI theory.

Chapter 4, *Quality of Assessment*, will explain quality of assessment issues, where the international concern of assessment is shown, followed by illustrations of some assessment debates throughout related literature; namely, criterion versus norm referencing tests, external versus internal (teacher) assessment, and higher cognitive demand (HCD) skills assessment. The role of teacher-based assessment in the KSA is then presented, in order to show why the present situation reveals that Saudi teachers

are, by far, not following good HCD testing practice. Since item analysis technique (IAT) is one of the core concepts of this research, its theoretical and mathematical ideas will be illustrated in more detail. Finally, the CAIAT software characteristics are presented in terms of how it facilitates the user running its functions, which helps to make it a user-friendly application and acts as one of its primary strengths.

Chapter 5, *Method and Design*, will explain that the method of this research will follow the descriptive method in considering the traditions of qualitative and quantitative inquiry. Since this research is an evaluative study, some light will be shed on this area. A comparison between qualitative and quantitative research approaches will also be provided, following which the main project design issues will be described in detail; namely, population, sample, generalisation and work plan scheme. Research questions will be presented in the form of main questions followed by sub-questions, which are distributed over two categories, or dimensions, around which the project activities move: specifically, *effectiveness* and *adoption*. Data collection instruments will be presented, and the quality of these instruments will be discussed in the light of the relevant literature. Statistical methods which are chosen to analyse the findings of the study will be illustrated in some detail wherever necessary. Finally, the limits and limitations of this research will be addressed, followed by comments reflecting on my personal experience in conducting the fieldwork of the research.

In Chapter 6, *Findings*, the descriptive statistics that reveal background information about the sample teachers will be presented; then, the research main outcomes will be presented by answering the research questions. Generally, the chapter provides quantitative findings presented through charts, curves and tables of data or statistical indicators, alongside some additional qualitative findings such as case study observations, interview results, general observation notes, and work team members' comments. I will show how some findings triangulate each other, which creates confidence in the results found.

Finally, in Chapter 7, *Discussion and Conclusions*, descriptive statistics findings will be utilised for exploring the impact of the teachers' background information on their pre-abilities. Then the chapter's main sections will follow, based on the study's dimensions – *effectiveness* and *adoption*. The inter-relationship between the different findings will be outlined. In addition, the connection between those and previous

similar research and related literature will be highlighted. The main conclusions and recommendations are articulated at the end of the chapter.

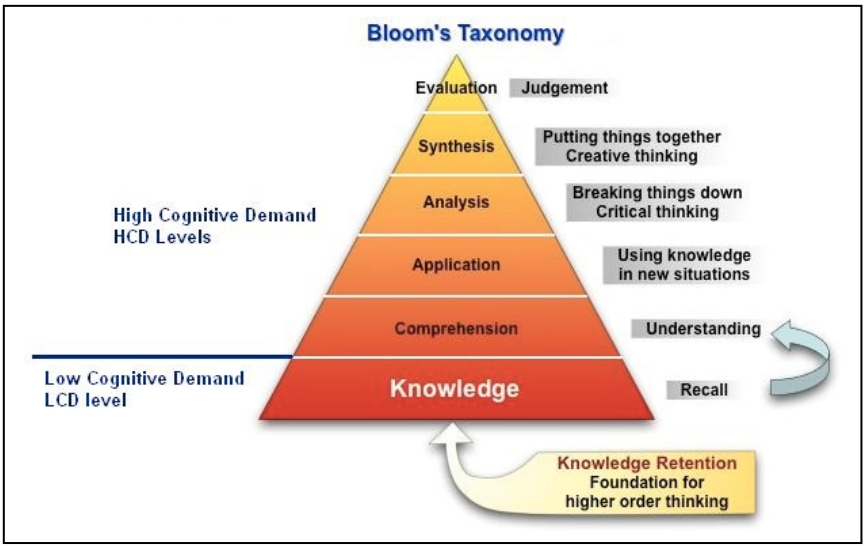
1.3 Main Concepts and Terminologies in the Research

The prime goal of the present chapter is to explain what this research is about, what problems it targets and in what way it will attempt to solve these issues. To provide a platform for all of these presentations, it is a fundamental requirement to outline the main concepts and terms underpinning and comprising the related language on this subject.

1.3.1 Bloom Taxonomy and HCD Questions

Bloom, Engelhart, Hill, Furst and Krathwohl (1985) categorised their taxonomy of cognition into six levels. As Figure 1.1 illustrates, they divided these into two main demands: low cognitive demand (LCD), which requires the mental skill of recall, and high cognitive demand (HCD), which includes the other five levels; namely, comprehension, application, analysis, synthesis, and evaluation[3] (Bloom et al., 1985, pp. 276-285).

Figure 1.1: Bloom Taxonomy of Cognitive Domain
(APU instructional material, 2009)



These five levels are widely considered in the literature of cognitive learning as key concepts of critical thinking. However, there is a disagreement about the classification of higher levels of Bloom's taxonomy. Stanley and Hopkins (1978, p. 179) have mentioned the difficulty to distinguish *application* items from those other higher levels citing many authors that reported the difficulty to agree on levels other than knowledge. Nevertheless, this does not mean that HCD levels are of little importance nowadays for either research or practice, as there are many educators who appreciate their concept and make use of their functions. Evidence of this is provided by the search that I carried out on the ERIC database using the keywords *Bloom's Taxonomy* for 1980 onwards, which produced a record of as many as 40 studies. Anderson (1999) presented her new vision of the taxonomy in her study *Rethinking Bloom's Taxonomy: Implications for Testing and Assessment*, which drew on Bloom's original work intensively and with a full appreciation of HCD levels. Interestingly, Stanley and Hopkins (1978, p. 179) stressed the importance of the taxonomy whereby "a teacher who has been exposed to the taxonomy […] can no longer be satisfied with a test that measures only rote learning of isolated facts." Aviles' (1999) experience revealed that:

> Creating tests using Bloom's taxonomy takes time and attention but the effort is well worth it […] I can also make my tests more challenging by teaching and testing to higher knowledge levels (Aviles, 1999).

1.3.2 IAT and CAIAT Concepts

Drawing on the classical theory of testing[4], IAT is one of the best methods for a thorough investigation of the quality of test items. Each question in a test is named 'an item,' and one uses the marks obtained in these items in a given test to calculate specific parameters for each item. These parameters are mainly the *difficulty coefficient* and *discrimination index*, whose names ably reveal their functions. They will be explained in more technical detail in Chapter 4; however, they are, in general, indicators from which one can infer the strengths or weaknesses of their corresponding item. Both parameters may indicate to the teacher that his or her question may be ambiguous, trivial, or undiscriminating, which alerts him/her to those items that are most likely invalid and can then either be edited, deleted or replaced with better alternatives.

Calculating IAT parameters, however, tends to be a difficult task because it implies the use of different equations with different types of variables[5]. In addition, dealing sometimes with a large number of marks increases the likelihood of errors. This difficulty, as well as the heavy workload it implies, discourages teachers, who are busy with many classes and different subjects, from using the technique for the sake of improving their tests. Wang, Wang and Huang (2008, p. 461) confirm that teachers may encounter problems when performing item analysis because "the statistics are complex and difficult to perform without the aid of a dedicated [digital] system." This digital system is a computer aided item analysis technique (CAIAT). The principle of the present research's CAIAT software package (which will be called 'the CAIAT' throughout the text) is that it is possible to overcome the obstacle of the statistical calculation part of IAT by means of computers. Davis (1975) considered that finding the discrimination index and difficulty coefficient by employing computers can be achieved easily and quickly, which leads finally to an increase in the validity of an examination (Davis, 1975, p. 74).

1.4 Statement of the Problem

Tackling HCD levels, during instruction and in assessment, is the bridge to higher quality education which encourages analytic skills at higher levels. The main problem that this research attempts to solve is the issue of the difficulty experienced by Saudi teachers in tackling HCD. This lack of success is seen as originating from two sources: the teachers' lack of HCD-level conceptual knowledge and pedagogical skills, and their lack of HCD test construction skills. As will be illustrated later, previous studies about Saudi education and some other investigations have revealed this perspective; one important reason for which, as indicated by these studies as well as attempts by the Saudi Ministry of Education (MoE) to effect improvement, is the lack of teachers' professional development (PD) pertaining to skills in test construction; especially those PD activities following appropriate scientific and normative approaches such as IAT. Although some attempts have been made in this respect (later on, will be explained further), these did not accomplish their aims because of difficulties accompanying the classical hand-calculation method used for IAT in those attempts, which is characterised by the substantial mental effort and time-consumption that discourage teachers from utilising IAT in the first place. As a consequence, another

approach is required for tackling this area of concern besides training the teachers in HCD knowledge and skills.

1.5 Aim

Many authors in the assessment field have recommended the utilisation of computers as this would be very likely to encourage teachers to practise item analysis and hence improve their tests (Robinson & Austin, 1969; Miller, 1970; Nitko & Hsu, 1984a, 1984b; Marso & Pigge, 1987; Wainer, 1989; Antelmo, Costagliola, Ferrucci and Fuccella, 2005; Bermundo & Bermundo, 2007; Manaligod, 2009). For example, Marso & Pigge (1987, p. 12) found that

> it would seem that most school systems need to increase the support available to assist teachers in meeting their testing and related responsibilities. This is particularly true in regard to clerical and computer support services; these would certainly appear to be essential to the improvement of teacher testing.

Also, Antelmo et al. (2005, p. 77) believe that

> It would be quite an interesting feature for a Computer Aided Assessment (CAA) tool to tell the tutor how her/his questions are effective in judging the learners. This information [item analysis] is a good feedback for the tutor, useful for modifying and improving her/his assessment material. Every time a question is modified, a newer and, hopefully, better version of it is generated and the lifecycle of the question continues.

Moreover, Manaligod (2009, p. 100) indicated the extent to which IAT have an impact on teachers' practices

> The valuable insights derived from the item analysis will enhance the delivery of educational content by way of effective and immediate feedback on the quality of the items and the scale as a whole.

Therefore, the aim of this research is to evaluate utilising the CAIAT computer software within what I have called the *HCD/CAIAT project* and find out whether it will succeed in stimulating teachers' PD in their assessment especially for the items of HCD level; specifically, *will teachers' abilities in constructing HCD questions and teaching at HCD level improve as a result of this project's intervention*?

Technically, this research assumes that the CAIAT software could encourage teachers to utilise the classical IAT in order to reveal (quickly and easily) the weaknesses and strengths of their test questions each time they administer a test. Thus, they would be able to improve their next versions of the same questions, and thus over time could improve their overall ability in test construction. This assumption is made on the basis of previous efforts, as shown in the 'Previous Studies' section that follows shortly.

Although a number of computer software packages do this job, they are mainly applied to multiple-choice questions and are in languages other than Arabic (mainly in English); thus, any software introduced to schools or teachers in the KSA should be redesigned in Arabic and cover all types of questions. Consequently, I designed the CAIAT software package in Arabic, to cover both types of questions – objective and essay – and applied it in the field within this evaluative research in order to examine its functionality (more technical elaboration of the CAIAT software can be seen in Section 2 of Appendix 1).

1.6 Literature Review

As the title implies, this research includes three important themes of education. The first is the *higher cognitive demand* (HCD) of Bloom's taxonomy, which expresses all related or similar concepts such as critical thinking, analytical reasoning, synthesis, problem-solving, and higher mental processes (Scriven & Paul, 1992 in Hummel & Huitt, 1994.) The existence of HCD in pedagogical practices is commonly seen as an indication of good instruction. The second theme is *teacher assessment* (TA) quality, not only in respect of examinations, but also as a formative assessment that is seen as a pivotal role for best practice in all other aspects of the teaching profession. The third theme is the *item analysis technique* (IAT) as a tool for teachers to improve their ability to write good test questions; and in particular, the test questions at HCD levels.

In the following sections, I will shed some light on these three concepts and a number of related issues, as revealed by the relevant literature and contextual implication. The main question targeted by this review is: in relation to quality of education, what level of importance are these three areas accorded in literature? Moreover, there will be a review of the Saudi local literature that aims to answer the

question: To what extent do these areas lack quality in the KSA? Both questions are geared to create a platform for the rationale of this work, which follows this review; however, there is a need for a further theoretical elaboration on these three areas, but this elaboration will be presented later in the following chapters, so for now they will be introduced by a more concise overview.

It is necessary to highlight from the outset that for the retrieved online sources, in this section and throughout the research, I was keen to use the most trusted sources for providing viable links such as ERIC, ProQuest, ScienceDirect, JSTOR, Questia and Wiley. In addition, I refer on occasion to sources other than these well-known ones, but made sure that indicators for the strength of these sources exist; for example, the availability of a '.gov' extension within the link, which indicates that the information is related to a governmental department (as in Albanese & Mitchell, 1993; Bermundo & Bermundo, 2007; EIA, 2001; Myron, 2005; Osborne, 2005), or the availability of the '.edu' extension, which indicates affiliation to an educational institution (as in Al-Agha, 2004; Dillon & Morris, 1996).

The reviewed local literature is mainly at Masters' level. My appraisal of a number of these works is that their aims tend to be simple and do not target in-depth problems or issues. However, most of my utilisation of these studies was in the form of making use of their findings that inform the essential question of rationality – *to what extent the problem exists* – which fits with the treatment level of these studies. Also, it should be noted that these studies are few in number, which is a reflection of two aspects of Saudi higher education research: a lack of originality in tackling serious issues, so many studies are repetitions of others' work but are merely applied to other states, provinces, or regions, and a lack of good archiving. As such, the researcher experiences a great deal of difficulty in obtaining information from Saudi sources, which is the main reason why the amount of cited local works seems scarce.

1.6.1 Higher Cognitive Demand (HCD)

Research indicates the positive impact of the HCD of Bloom's taxonomy on students' learning and creativity. In their study *Critical Thinking in the Management Classroom,* Athanassiou, McNett and Harvey (2003) found that using Bloom's taxonomy as a supporting device moves students toward self-management in learning,

while repeated emphasis of Bloom's concepts improves higher-level conceptual thinking. Sultana (2001) pointed out that HCD levels in Bloom's taxonomy are classified as thinking skills levels that reflect the mental processes of conceiving, manipulating and dealing abstractly with ideas. His research data analysis revealed that only 23% of the objectives written by American first year teachers in their lesson plans were designed to develop HCD skills. Earlier, though, Cole & Williams (1973) studied the correlation between the cognitive level of teachers' questions and the corresponding cognitive level of their pupils' answers and found that it is significant. In her study on reducing the gap between inferential and literal comprehension[6] for 105 second graders, Coolidge (1989) found that after training teachers to use HCD questions in learning activities and providing opportunities for students to practise answering questions, this gap decreased from 37% to 13%.

The quality of teacher questioning during instruction has a relationship to HCD it terms of facilitating discussion and guiding students' efforts at acquiring a cognitive or social skills (Gage and Berliner, 1984, pp. 632-633). Cotton (1988, p. 3) reported in a systematic review of related research that "oral questions posed during classroom recitations are more effective in fostering learning than are written questions." More elaboration on questioning will come during the 'Assessing HCD Skills' section in Chapter 4; however, the importance of questioning at HCD levels, as illustrated above, is the reason behind this research focusing on HCD in order to improve teachers' ability to write (and hence ask verbally) optimal HCD questions.

1.6.2 HCD in KSA

Ayedh (1993) studied the extent to which secondary boys' school physics objectives were achieved, from the perspective of Riyadh teachers and educational supervisors, and found that the aims of the cognitive domain are fulfilled relatively at a 76.7% success rate. Among obstacles that hinder the optimum fulfilment of these aims (in other words, much higher that 76.7%) are high teacher workloads and the low level of quality of physics textbooks in terms of the way they explain and demonstrate scientific ideas and concepts. Furthermore, he found that HCD instructional objectives were being achieved less than LCD objectives were, which he attributed to a number of reasons. The first was the way of presenting the subject in textbooks, which gives much more attention to LCD objectives than HCD objectives. The second is teacher emphasis

on LCD objectives during instruction, while the third is the focus of teachers' tests on LCD levels, which invites pupils to pay attention to this level of objectives during their learning (Ayedh, 1993, pp. 80-84). In his study, Al-Saif (1981) recommended guidelines for the science education programme in Saudi state secondary schools, based on the levels of thinking of Saudi secondary students compared to their US counterparts. The findings revealed that science teachers lacked knowledge about pupil learning and teaching methods and that the majority of science teachers had not considered in-service education (Al-Saif, 1981, pp. 160-164).

Al-Darweesh (1999) studied the extent to which science teaching's aims are achieved in elementary boys' schooling from the perspective of the science teachers in the province of Al-Kharj[7]. His findings revealed that LCD aims are dominant compared to HCD aims and that obstacles that affected optimum fulfilment of teaching aims represented a percentage of 79%, which is considered significant. These obstacles were mainly a high teaching workload (which is represented by weekly lesson plans of 24 lessons), a lack of instructional aids and facilities, and high pupil class numbers (Al-Darweesh, 1999, pp. 86-87).

Besides the issue of the deficiencies in teachers' professional abilities to instruct at HCD level, curricular textbooks do not support or subsidise this point. In his survey of a number of science teaching problems in intermediate schooling in Riyadh, KSA, Abdulmuttaleb (1984) unearthed a number of findings regarding pupils' textbooks. Among these was that they did not promote self-learning, nor set out questions that provoked pupils' thinking, nor provide comprehensive, evaluative questions in any great number (Abdulmuttaleb, 1984, pp. 249-259). It seems that this trend affects a number of Arab states. Several Arab researchers have conducted their studies using Bloom's taxonomy for analysing textbooks questions. Table 1.1 summarises their findings (Al-Agha, 2004, pp. 456-457, p. 462).

Table 1.1: Summary of findings from a selection of Arab studies

Author	Year	Subject	Percentage of LCD questions	Percentage of HCD questions	Main HCD levels
Abo Al'ola and Abdulhameed	1985		64%	26%	Comprehension
Al-Agha	1994	Science, year 9 in Palestine	95%	5%	-
Fadhel	1998	History, year 6 in Iraq	71.4%	24.4%	Comprehension
Zaytoon	1999	Science, year 9 in Jordan	47.6%	45.3%	Comprehension (34.4%) and Application (10.9%)
Al-Agha	2004	Geography, year 6 in Palestine	79.33%	20.67%	Comprehension (14.67%), Application (3.33%), Analysis (2%), and Synthesis (0.67%)

I have analysed the content of a number of teacher-written tests to discover to what extent this problem exists. From secondary school physics teachers in the city of Al-Ahsa, KSA, I randomly selected 21 written examination papers for the end of the first term of the academic year 2003/2004. As Table 1.2 shows, the findings indicate very low interest in tackling HCD. I denote that the value 23% as representing the mathematical problems solving should not be looked at as high because this area represent most of the subjects' curricular content.

Table 1.2: Distribution of levels of cognition across Al-Ahsa physics teacher-written tests

Main Level	Level of Cognition	Number of Items	Percentage
HCD	Synthesis	1	0.2 %
	Analysis	1	0.2 %
	Application (of concepts)	29	4.5 %
	Application (in solving mathematical physics problems)	149	23 %
	Comprehension	80	12 %
LCD	Recall	392	60 %
	Total	652	99.9 %

Moreover, I read the Al-Ahsa science educational supervisors' annual reports about their teachers' tests and found that most of them commented on the lack of HCD questions in those tests. I briefly interviewed some of them regarding this issue, and found that they attribute this paucity to three main reasons: first, the lack of teacher competency in writing such questions; second, the lack of national tests as a bench mark; and third, the lack of authentic teacher performance evaluation, along with the presence of automatic annual increments in teachers' salaries, regardless of their levels of performance. The first reason is in line with what Al-Bakr (1998) found in her study on physics test questions for final year secondary school pupils, where she identified a low incidence of HCD questions and hence recommended that test writers need to be trained. The third reason is in line with what most of the visited studies have revealed about weaknesses in teachers' performance, both in instruction and question construction. In addition, what Al-Dakheel (1997) established about the contribution of the *employee performance evaluation form* in raising the quality of teachers' work may give this finding further credence. Al-Dakheel (1997) explored the efficacy of the *employee performance evaluation form* from the perspective of Riyadh principals and educational supervisors. Among her findings was a paucity of follow-up procedures regarding performance evaluation results. She also found that 32% of the respondents did not believe that this form contributed to diagnosing teacher performance

weaknesses. Their opinions about the functionality of the form in estimating competency, achieving educational aims, and revealing training needs were in percentages of a similar range. I think that this level of opinion needs to be considered when looking at the important contribution of this form to the quality of education, as highlighted by educational supervisors' annual reports and other documentation.

1.6.3 Teacher-based Assessment (TA)

The quality of any teacher's instruction could be defined through the level of his or her questioning. In many cases, the statement within the question reveals the validity and/or functionality of the question. Different factors work in this respect, for example the use of words, using variables, any different conditions, circumstances, limits, or any special considerations, should all be made clear so that the question lies within the context about which it is asked. Dealing with these factors becomes much more difficult when the question tackles an HCD level.

By evaluating 95 elementary and secondary school teachers' knowledge of testing and measurement, Daniel & King (1998, p. 331) found that "the teachers' knowledge bases were somewhat inadequate." Pupils infer teachers' aims and attitudes from trends in their test questions. This highlights the argument that improving the teaching-learning process cannot succeed through classic testing practices in which most of the emphasis (if not all) is on information recall (Baez, 1970, pp. 273-275). Pragmatically, "if any part of teaching is to be assessed, then it should be the 'business' of quality control of assessment" (Clift and Imrie, 1981, p. 129). Adkins (1974) uncovered some deficiencies in teacher-made tests and pointed out their side-effects in the learning atmosphere:

> Individual students often find serious fault with teacher-made tests – ambiguities in questions or alternatives […] Some students do not complain openly, and such situations rankle with them and may adversely affect motivation (Adkins, 1974, p. 132).

The way Adkins expressed how poor assessment reflects negatively on pupils' learning applies to any education situation, since it subscribes to the psychology of learning and social pupil-teacher and pupil-pupil interactions. It is suggested that teachers should not underestimate the difficulty of framing good test questions where "each test item is read

very carefully by most examinees: each word becomes important, and the examinee is constantly asking what the item or word *really* means" (Bloom et al. 1981, p. 181).

1.6.4 TA in the KSA

Zahranee (1998) studied the extent to which intermediate school science teachers in Makkah, KSA, are competent in test construction skills, as well as their level of practice in this respect, and found low levels in both. He recommended that assessment courses at colleges of education should be redesigned to meet actual field requirements, so instead of focusing on developing textbook content, teacher PRESET and INSET should be given attention and teachers should be trained in how to use computers in order to build up test banks according to modern models of assessment theory such as the Rash model, which is an IRT[8] (Item Response Theory) model. In his study, Hejran (2002) explored training needs for Saudi teachers from the viewpoints of educational leaders, professionals and educational supervisors. Depending on the sample individuals' perspectives (110 individuals), he sorted 50 training needs and found that the teachers' need for training in assessment came in first place.

1.6.5 Item Analysis Technique (IAT) and CAIAT

Item analysis procedures contribute greatly to increasing the reliability of a test, and subsequently to improving the test's potential validity[9]. The reliability of a test is statistically dependent on the standard deviation of the same test, which in turn is "related to the level of discrimination of the items," with consideration for optimum difficulty at the same time (Hills, 1981, p. 74). Coniam's (2009) research examined, in a quasi-experimental study, the quality of tests that Hong Kong English (EFL) teachers produced for their students. They were asked to construct objective tests and implement item analysis, and then to reflect on their experiences about the test development process and to examine their test quality using IAT. In general, the number of 'good' items was lower than their expectations. They commented that "the tests they had previously produced [before the study and during their work] had not provided them with usable, accurate information about their students' abilities."

I have not encountered opinions against using IAT for improving teachers' testing abilities, and it seems that there is consensus in this respect because many researchers and assessment experts recommend IAT as a necessary method for

improving testing practices. Collins, Johansen and Johnson (1976) explained how teacher-made tests could be improved through four steps: item analysis, computing validity, computing reliability and going over the test with pupils (Collins et al., 1976, pp. 98-107). They posited that going over the test with pupils is one of the best opportunities for the teacher to clear up misunderstandings, provide remedial help, and reinforce learning (Collins et al., 1976: 106). Moreover, Nitko (1983) considered item analysis practice as being an important source of feedback for the teaching-learning process, as it provides feedback not only to pupils about their performance, but also to teachers about pupils' difficulties and about areas of weakness in the curriculum (Nitko, 1983, p. 298).

Since the early 1980s, computer-based IAT has been widely utilised by institutions and researchers, not only for research or psychometric purposes but also for helping instructors and teachers to improve their assessment. Carlton University in Canada, for instance, used ESP, a kind of CAIAT software, written by Professor Alan Moffitt of the Psychology Department. The university's website states its aim as follows: "New in 1995 is item analysis software developed to assist instructors to improve the quality of their multiple-choice questions" (WTLRCCU, 2002). Kim (1999), as a CAIAT software developer and an educationalist specialised in this field, indicated the important role of computer technology and software development in making IAT feasible via user-friendly software packages. Yu (2006, p. 1) considered that classical IAT is frequently used by teachers for its "conceptual and computational simplicity" and suggested ways of utilising the famous statistical software application SAS© in this respect.

1.6.6 IAT in KSA

One of the main functions of the General Directorate for Educational Measurements and Evaluation (GDEME)[10] at the Saudi MoE is to carry out field studies and discover shortcomings in assessment processes in schools. It is also required to develop pupil assessment techniques and to take care of evaluation issues within different educational settings. The GDEME has been given the responsibility of supervising the 'Achievement Tests Project,' which started in 1999 and includes the production of item banks for every subject taught from year 4 (fourth year of primary school) to year 11 (second year of secondary school). To broaden my insight into the

project, I met the founders and the people in charge of this project and asked them a variety of questions about their rationale for the project, their experiences so far, and what their expectations were for the project. I found that they initiated the scheme because of poor teacher-made test quality. Interestingly, there are no national or standardised tests that can be utilised as frames of reference to judge teacher-made tests, but the project founders made use of comparison studies available at the Ministry, which showed correlations between teacher-made test results in the first semester of the third year of secondary school and their graduation test, which is the national test at the end of the second semester. The latter is designed by the Ministry and thought to be a more 'objective' or at least an 'unbiased' test that can be used as a reference for comparison. The findings revealed that teacher-made tests needed reform.

As a result, two projects have been initiated. The first is the items bank project and the second is a training scheme for test construction and IAT. The items bank is to be distributed to schools all over the kingdom to replace teacher-made tests. The MoE's aim by effecting this initiative is to eliminate errors and weaknesses that appear in teachers' written questions[11]. The second project (the training scheme) includes many courses that have been held around the kingdom for teachers, principals and supervisors on test construction, accompanied by an introduction to IAT. The purpose of these courses has been to raise awareness among teachers about constructing valid test questions and to help principals and supervisors to use test results in evaluating, training and supervising their teachers. Being closely involved in some of these programmes, I noticed that the participants often felt frustrated when they reached the statistical part of the course, and they indicated that difficult calculations (not computerised) and the amount of time needed would probably prevent them from implementing what they had learnt. Furthermore, the statistical part of the course inevitably takes up most of the training time, and hence shortens the training in other skills and any discussion time available for understanding item analysis; that is, the functions of its parameters.

1.7 Rationale

The efforts of the GDEME mentioned above reflect the level of the Saudi MoE's awareness of the problem regarding low quality TA, which is represented mainly by the absence of HCD questions. Local Saudi research confirms the existence of shortcomings in TA, especially in relation to HCD levels, which were revealed as early

as 1975 (Al-Mazyed, 1975). The MoE's Achievement Tests Project aimed to provide teachers with an items bank that overcomes the lack of professional preparation in constructing good questions, as Al-Mazyed's (1975) research found. This was cross-validated by Al-Zahranee (1998), who reported that science teachers had poor test construction skills, so he recommended the items bank approach. Furthermore, other studies (Al-Saif, 1982; Ayedh, 1993; Al-Darweesh, 1999) consolidated this perspective whereby they found that the teachers' skills or practices regarding HCD, and levels of understanding by students, were low.

The MoE's training scheme for test construction and IAT is based on the MoE's observations that teachers make errors in their test questions and do not tackle HCD levels. This is also confirmed by the content analysis I carried out on a sample of teachers' tests on which few HCD questions were found. My review of educational supervisors' annual reports revealed similar findings which also in line with Tashkandi's (1981) research findings in that it determined that training pre-service teachers in this respect contributed to increasing their corresponding skill and practice. The shared aspect for these different efforts is the urgent need to train teachers. In addition, Al-Saif (1981) and Ayedh (1993) pointed out this aspect and shared Zahranee (1998) and Hejran's (2002) recommendations for training teachers in these skills. Specifically, Zahranee (1998) and Hejran (2002) stressed the need to train teachers on test item construction and IAT. All of these contribute to strengthening the rationale of the present project's trend.

The local studies that I have cited above, dating from 1975 to 2004 and covering different key stages, may give an idea about former and current situations, as they highlight a number of shortcomings and outline major obstacles. Moreover, from my experience as a former science teacher and educational supervisor in the KSA for about 15 years, and as a commentary conclusion on what has been illustrated, I see that these problems are significant and a huge effort will be required to overcome them, in order to provide a better educational system. Furthermore, the studies indicated the low level of interest in HCD, low rate of existence of HCD questions in school or national tests, and low teacher competency in constructing HCD questions. Therefore, these studies called for teacher training so the shortcomings could be overcome. The present project, therefore, is considered to be helping to fulfil this need in terms of the quality of

assessment and specialised teacher training, as it represents a response to what the local studies in the KSA revealed and integrates the training effort of the GDEME to improve teachers' abilities in IAT. Essentially, it provides them with a new methodology that triggers their self-learning skills and fills the gap in teacher assessment literacy.

1.8 Previous Studies

It seems that little research has been conducted in this area. In his research about assessment and reporting arrangements facilitated by a computerised record-keeping system in a UK secondary school, Comley (2001) reported that one of the difficulties he experienced was finding existing research on the subject. If this is so in an administrative area of this kind, then it is no wonder that the present research faces the same situation, if not worse. To the best of my knowledge, no similar effort has been undertaken in the KSA, so all illustrated previous studies are non-local ones, except for Tashkandi (1981), which falls in line with the present research in one aspect: the HCD dimension. The main question that this presentation aims to answer, therefore, is: What other similar works exist, in terms of conceptual arguments included and practical lessons learnt, that can inform the present work?

Weiss (2011) recently illustrated the historical development of utilising computers in educational testing. He highlighted early instances from the 1960s, involving scanning and item analysis, and showed that the emergence of microcomputers in the mid-1970s expanded these first steps even further. Some writers, such as Kumar (1997), have presented a number of uses of computers in assessment, which shows the extent to which this area is rich and research-appealing. My initial observation is that the majority of research in this area is dedicated to purely psychometric science and focuses extensively on modern IAT trends such as IRT, which benefits those working in higher levels of assessment, not at school level. Sadly, less attention is given to teachers' practices in IAT or to improving their ability in this respect. Even studies that did tackle test item analysis aided by a computer did not look at the impact on teachers as much as they focused on technically describing the electronic system they created and how it could be utilised by teachers or schools (for instance, Wainer, 1989; Kim, 1999; Antelmo et al., 2005; Yu, 2006; Manaligod, 2009; Sukamolson, undated). In some cases the creation or assembly of the test was the goal of the software (for example, Millman & Westman, 1989; Mitkov and Ha,

2003; Brown, Frishkoff, and Eskenazi, 2005; Linden & Diao, 2011). Another thread followed by quite a large number of studies is that of tackling the creation of test *item banks* to help teachers assemble tests and in some cases generate a test form (for instance, Wright & Bell, 1984; Fairon, 1999; Eggen & Straetmans, 2000; Maughan & Willmott, 2001; Anzaldua, 2002; Weiss, 2011; Yang, Han, and Zhou, 2011). The widespread approach of computer-adapted tests (CATs) is a major component of the last thread; nevertheless, depending on ready-made item banks, in my opinion, is less constructive for classroom purposes than empowering teachers to ask optimal questions by increasing their ability to write good test questions by means of on-going IAT good practice, which it is suggested the CAIAT will invoke. In their critical review of selected computer-assisted language testing instruments, Silye and Wiwczarosk (2002) commented that:

> It should be clear, though, that neither the [computer-based testing (CBT)] schemes nor the [web-based test (WBT)] themselves […] make a good language test without sophisticated expert knowledge of test writing and validation provided.

Although it might be considered that by using item banks teachers save time which can then be spent on teaching and improving instruction rather than developing test materials (Pearson Inc., 2005), I believe that teachers who depend continually on ready-made test item banks are losing the opportunity to analyse their tests and thus improve their ability to construct optimal questions. As such, this is very likely to have a negative impact on the quality of instruction because a teacher's ability to write good test questions is very likely to determine his/her ability to ask good questions verbally during instruction. Actually, questioning is a core dimension of constructivist pedagogical practices in the classroom, in its role as a tool that provokes students' thinking. John Dewey said: "What's in a question, you ask? Everything… It is, in essence, the very core of teaching" (Cotton, 2011). Item banks are better utilised by teachers as a database that hosts tests and as an instrument that analyses and/or validates those tests, which maintains the on-going practice of constructing optimal tests by teachers. Since a great deal of the literature covers the use of ICT in assessment, it might be noted that this review has, to some extent, shifted to the area of teacher test quality from a psychometric perspective.

An early attempt to tackle the utilisation of IAT concept was undertaken by Smawley (1962, p. 138) through his suggested *Shortcut IAT*, which includes two quick methods for item analysis that he considered would "suffice for most teachers' purposes in the classroom." This attempt does not include the utilisation of computers but instead provides teachers with an easy calculating method as opposed to the classical one that requires considerable mental effort and work time. This is an early indication of the importance of providing teachers with an easy-to-use instrument for IAT, which lies at the heart of the present work's rationale.

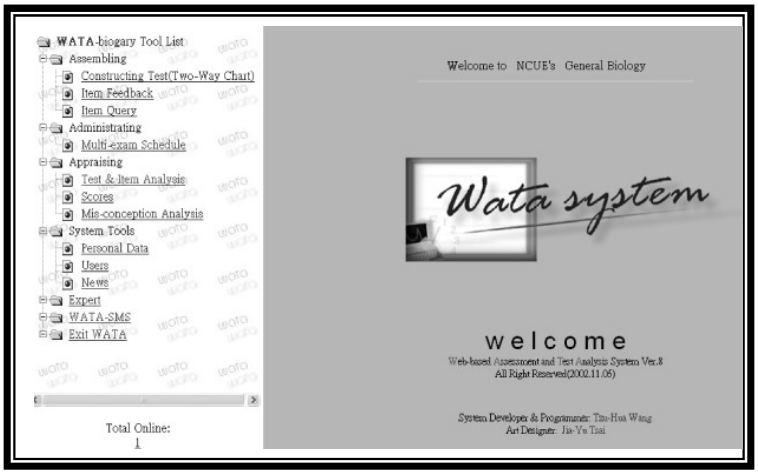
In Grossmont District, USA, Robinson & Austin (1969) conducted a study into the use of a "computerized test-correcting service for teachers," in which they applied a computerised system in a school to provide teachers with statistics about his/her tests, as well as item analysis parameters. Their observations included saving a huge amount of teachers' time ordinarily allotted for processing and grading tests. For the school year 1967-1968, they estimated an actual saving of at least 100 teaching hours; moreover, they indicated that the resulting statistical data "surpasses" those that were paper- or pencil-produced. They also indicated the successful case of a teacher who noted that a student's results in *Applied Arts World History* were two standard deviations above his class mean, which prompted the teacher to transfer the student to a higher class. That student was able to adjust very well and continued to score high marks, and this was considered by the researcher to be a positive outcome from using a computer-aided methodology. Nevertheless, I believe that this case is not necessarily an indicator of the benefit of using CATs exclusively but of good teacher training in dealing with the outputs of these analyses. In the present research, the PD of the teacher, through training and the on-going practice of self-learning, which the CAIAT software triggers and expedites, is considered a pillar of the project and parallel to the implementation of the new computer system.

Wayne (1976) wrote about the efforts of Beaverton School, in the USA, which adopted a model to help teachers with the construction and analysis of diagnostic tests. The model consisted of three services: IAT by computer, test scoring and summarisation, and data storage and retrieval. He reported that "the effects of the data analysis programs in three years of a 9th grade writing course showed some evidence of student improvement in meeting course objectives."

Nitko & Hsu (1984b) also suggested a computerised IAT for teachers. They have reviewed approximately 50 item statistics[12] in order to determine the most appropriate one to be incorporated into a computerised IAT system for classroom teacher (Nitko & Hsu, 1984a). By utilising a Z-test calculation in order to compare the elapsed time of two methods of item analysis – with a computer and without (that is, CAIAT and IAT) – Bermundo & Bermundo (2007) found that the CAIAT time was significantly less than that of manual IAT, which outlines the high feasibility of introducing computer-aided IAT to teachers.

Wang, Wang, Wang, Huang, and Chen (2004) developed an assessment system for a Web-based test analysis program called the WATA (Figure 1.2). One of the WATA's system functions was to "generate test results and analyses for teachers." To examine its impact on teacher education, Wang et al. (2004) applied two studies. The first was conducted in 2001, during the development of the system, with a sample of 47 in-service teachers who were asked to assess the functions of the WATA. The findings indicated a good level of satisfaction. The second study, which lasted for four months, involved 30 pre-service teachers using the WATA system during a teacher training program. The findings revealed that their perspectives on assessment changed significantly (Wang et al., 2004).

Figure 1.2: A sample screenshot of the WATA (Wang et al., 2004)



As Table 1.3 shows, using the WATA system significantly transformed participants' perspectives about classroom assessment in the five areas highlighted by the ** sign. Among these, three relate to the present research directly; namely, "to understand the

strengths and weaknesses of test items," "to understand the strengths and weaknesses of choices for a multiple-choice item," and "to understand the strengths and weaknesses of a test." Furthermore, the participants' use of the WATA system made them more willing to follow the standard procedures involved in administering a test. As Table 1.4 shows, there were "significant transformations in all steps of assessment except for the three steps required for every test: 'assembling the test', 'administering the test', and 'scoring the test'." I consider the WATA system a good similar example of the present project's CAIAT system.

Table 1.3: The impact of the WATA system on teachers' perspectives about the purpose of assessment (Wang et al., 2004)

Table 3. Changes in perspectives of the participants on purpose of assessment by using the WATA system ($n = 30$)

Purpose of assessment	Before WATA training (%)*	After WATA training (%)†	Cochran's Q test statistics
To distinguish learning outcomes of students	73.33	86.67	2.00
To understand students' popular confused conceptions (or misconceptions) about subject matter	70.00	76.67	0.40
To analyse an individual student's misconceptions	23.33	70.00	10.89**
To analyse and improve a teacher's instructional strategy	10.00	63.33	14.22**
To realise strength and weakness of test items	20.00	76.67	13.76**
To realise strength and weakness on choices of a multiple-choice item	16.67	90.00	22.00**
To realise strength and weakness of a test	26.67	80.00	12.80**

*Cronbach α: 0.65; †Cronbach α: 0.89; **$P<0.01$.

Table 1.4: The impact of the WATA system on teachers' perspectives about the assessment steps (Wang et al., 2004)

Table 4. Changes in perspectives of the participants on steps of assessment by using the WATA system ($n = 30$)

Basic steps of assessment	Before WATA training (%)*	After WATA training (%)†	Cochran's Q test statistics
Determining the purpose of testing	60.00	93.33	11.00**
Constructing the two-way chart	23.33	96.67	21.00**
Selecting appropriate items according to the two-way chart	23.33	93.33	22.00**
Preparing relevant items	60.00	100	11.00**
Assembling the test	100	100	N/A
Administering the test	100	100	N/A
Scoring the test	100	100	N/A
Appraising the test	23.33	100	23.00**

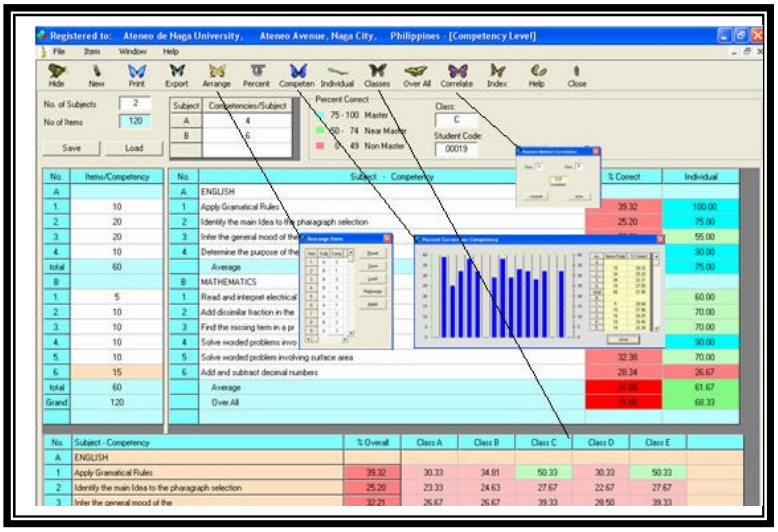
*Cronbach α: 0.73; †Cronbach α: 0.92; **$P<0.01$.
N/A = not available.

The findings of another recent experimental research paper by Wang et al. (2008) on two groups of pre-service biology teachers in Taiwan revealed that using the WATA system as a framework for teacher assessment literacy can effectively improve a teacher's assessment knowledge and perspective more than other non-Web-based frameworks do. The authors believed that those who use the WATA have the opportunity to "exploit" the system's statistical indicators when undertaking test item analysis and revising tests. The study further suggests that the WATA system should be considered a model for any assessment literacy development program for pre- and in-service teachers.

In addition, Fan, Wang and Wang (2011) investigated the effect of the WATA system on a sample of 47 secondary in-service mathematics and science teachers on a summer training programme. The research collected pre-test and post-test data about the sample teachers' knowledge and perspectives on the assessment process. The findings revealed that there was a significant improvement in teachers' assessment perspectives and that their knowledge was improved by training, especially for teachers with low prior knowledge.

Bermundo & Bermundo (2007) executed a project that included developing a form of CAIAT software which they called the "Test Checker and Item Analyzer with Statistics (TCIAS)" (Figure 1.3) and using it with 100 teachers, who were randomly selected through purposive sampling from four sectors in Camarines Sur, Philippines. Questionnaires, interviews and observations were applied to these teachers. The researchers' findings revealed that, on a scale of 1-5, the ratings given to the TCIAS by the teachers were 4.56 for usability and 4.35 for acceptability. The authors concluded that the TCIAS was perceived by the teachers to be "most useful and most acceptable."

Figure 1.3: A sample screenshot of the TCIAS (Bermundo & Bermundo, 2007)



Costagliola & Fuccella (2009) created a computer-aided IAT that they called the *Rule Based E-testing System*. It utilises a previous system called *eWorkbook*, which they created previously for use in the classroom (see Figure 1.4). In addition to providing the teacher/tutor with parameters for test item quality, such as difficulty and discrimination coefficients, it provides them with advice driven by what these parameters reveal. The system uses the traffic light colour system; in other words, green to show good items that need to be re-used, red for clearly weak items that need to be discarded, and amber for weak items that need to be modified or maybe discarded, with textual advice that highlights to the user the potential cause of weakness. The application implies that the teacher writes the test on the system and then it is applied online to the students. After the system suggests to the tutor what action needs to be undertaken for each item (as explained above), and the teacher improve those items that require modifying, it is assumed that they provide better test quality parameters the next time they are used in a subsequent test. Costagliola & Fuccella (2009, p. 5) explained that "the items can be evaluated by the system through subsequent test sessions, following the lifecycle shown in [Figure 1.5]." In general, quality improvement is obtained in two ways:

- Through the increment of item *discrimination*. This objective is pursued by both eliminating and opportunely modifying items with low *discrimination*.

- By having the tutor's estimation of the *difficulty* closer to the calculated difficulty for the item.

I believe that this way of tracking test items is a creative way to gain hands-on findings on the functionality of a system relating to teacher performance. However, I should emphasise that this was possible for that group of researchers because they were able to design the program as an 'online' system, whilst the programming part of the present project is limited to a computer-based system and has no capacity for online application.

Figure 1.4: A sample screenshot of the *eWorkbook* (Costagliola et al., 2007)

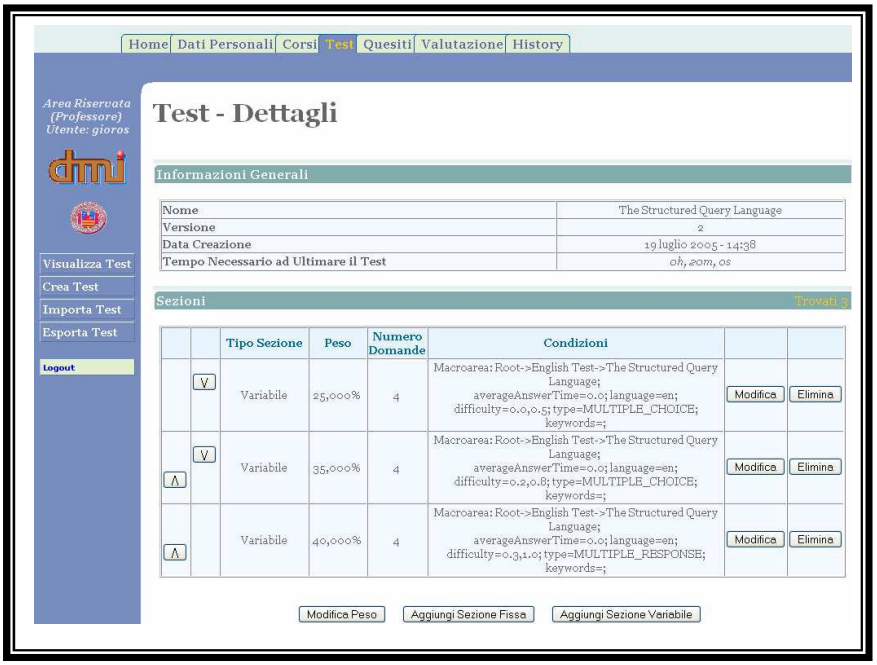
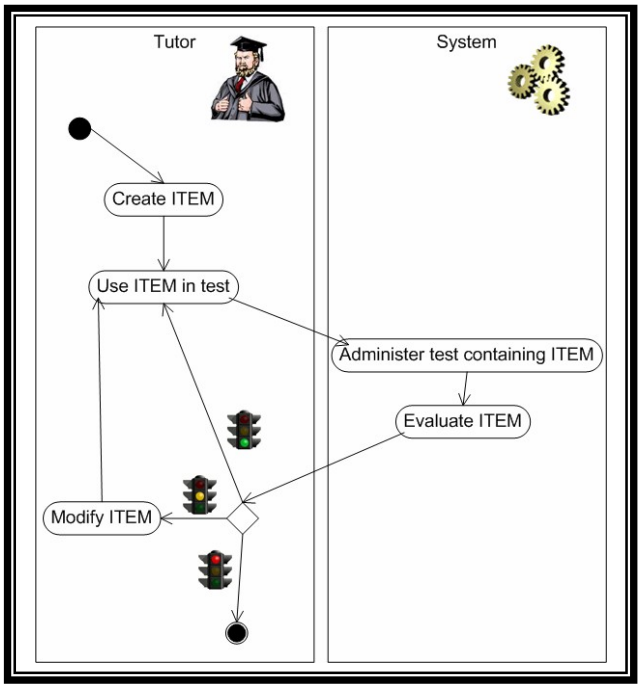


Figure 1.5: Item Lifecycle (Costagliola & Fuccella, 2009, p. 6)



Costagliola & Fuccella's system was tested on tutors on a course at the University of Salerno, and took place over two test sessions. After the first test session, the system highlighted which items needed to be discarded and which should be modified. After the required actions had been taken and the test re-administered with the new modified/updated version and to other students, the second test session revealed an improvement in the discrimination index from a value of 0.375, obtained in the first session, to 0.466. As for the difficulty parameter, there was a decrease in the mean difference between the difficulty estimated by the tutor and that calculated by the system, which was considered a positive signal of the system's functionality (Costagliola & Fuccella, 2009). Despite the fact that this research was carried out not by educationalists but by computer science researchers, whose aim was to apply a computer programming method known as *Fuzzy Classification*, their administration of the system considered the ease of use for any level of user and not merely for those specialised in programming methods. Their findings, as explained above, indicate that the tutors were able to use the system without difficulty and that they achieved significant outcomes in terms of what the system was designed for; that is, improving the quality parameters of test items. They reported that the system "produced encouraging results, showing that the system can effectively help the tutors to obtain items which better discriminate between skilled and untrained students and better match the difficulty estimated by the tutor" (Costagliola & Fuccella, 2009, p. 16). In another similar project using the same system, *eWorkbook*, Costagliola et al. (2007, p. 15) reported that "the testing has shown that teachers, also with very little technical skill, can easily use *eWorkbook* to create assessment tests."

Another theme of this research is HCD. Tashkandi (1981), a Saudi researcher, investigated the effects of training Saudi pre-service teachers in questioning techniques in order to produce educators who would use significantly more HCD during their teaching practice (Tashkandi, 1981, p. 61). He developed a 40-item "Question Classification Test QCT," based on Bloom's taxonomy of the cognitive domain, and administered it as pre- and post-tests to 45 pre-service science teachers (Tashkandi, 1981, p. 8). By following an experimental design of inquiry, the experimental group received training in questioning via a self-instruction program[13] and then were asked to teach in schools for three weeks. He found that the utilisation of a practice-oriented training program in question classification and techniques had a significant, positive

impact on the teachers' questioning practices, resulting in an increased number of HCD questions during instruction (Tashkandi, 1981, pp. 67-68).

To summarise, the previous studies have shown that tackling what this research aims to study is attainable, feasible and promising, which encourages the present project to invest in researching this area and building on what the previous works have found. In short, the major ideas this review has revealed are as follows:

- Little research has been conducted in this area.
- Historically, the earliest instances are from the 1960s (Smawley, 1962; Robinson & Austin, 1969; Wayne, 1976; Weiss, 2011).
- The majority of related research is dedicated to purely psychometric science of IAT with less attention devoted to teachers' practices.
- Studies that did tackle CAIAT focused on
    (a) technically describing the electronic system created,
    (b) the creation or assembly of the test, or
    (c) the creation of test *item banks* to help teachers assemble tests.
- Using a CAIAT resulted in saving a huge amount of teachers' time (Robinson & Austin, 1969), and some evidence of student improvement in meeting course objectives (Wayne, 1976)
- Nitko & Hsu (1984b) determined the most appropriate item statistics to be incorporated into a computerised IAT system for classroom teacher.
- The measured time for CAIAT was reported as significantly less than that of IAT (Bermundo & Bermundo, 2007).
- Teachers' perspectives about their satisfaction, and the ease of use, usability and acceptability in utilising a CAIAT system were positive (Wang et al., 2004; Bermundo & Bermundo, 2007; Costagliola & Fuccella, 2009).
- Applying an online CAIAT resulted in teachers' perspectives on assessment changing significantly, especially those about TA (Wang et al., 2004; Fan et al., 2011).
- Teachers utilising a CAIAT became more willing to follow the standard procedures involved in administering a test (Wang et al., 2004).
- Teachers' knowledge about IAT was improved by training, especially for teachers with low prior knowledge (Fan et al., 2011).

- There is a reported improvement in the parameters of discrimination and difficulty in teachers' tests as a result of their use of a CAIAT system (Costagliola & Fuccella, 2009).
- Training Saudi teachers in questioning techniques based on Bloom's taxonomy resulted in those teachers using significantly more HCD during their teaching practice (Tashkandi, 1981).

1.9 Significance

At the initiation stage of the present research, it was suggested that it should be linked to a national project for developing testing practices in Saudi schools, under the supervision of the former Vice-Minister for Girls' Education[14]. As such, this outlines its distinctive role as a reform tool that could be adopted by the MoE. Furthermore, and to the best of my knowledge, no similar studies have been carried out in the KSA and the Middle East, and it therefore represents an important addition to the literature. The CAIAT software proposed in the present study characterises a new and distinctive product that can be used by all Arab countries, as, again to the best of my knowledge, there is no similar Arabic software currently available for the teacher level[15]. Its merit is not confined to teacher-made testing, however, as it can also apply to different areas of interest such as research and standardised testing (under the classic test theory). Moreover, implementation procedures are likely to supply schools and science teachers in the research sample with a very rich experience in terms of computer uses for teaching, test item specification, HCD, IAT and the CAIAT. This research has provided me with genuine expertise in the way that an item analysis approach can be applied at the school level; that is to say, its implications, difficulties and effects. In addition, this enriched my understanding of item analysis parameters from a statistical perspective.

Further experimental studies could be developed from the present research to investigate, longitudinally or otherwise, the tangible effects of the CAIAT on such aspects as teachers' overall PD, testing-culture through schools, pupils' learning and parents' awareness of test findings. Because it includes decision-making by the teachers, this research informs other future studies that could cover the use of computer technology in schools for educational decision-making. Based on the present research findings, follow-up-studies could focus on the relationship between teacher-made HCD assessments and pupils' learning from other various dimensions.

Above all, I expect the influence of the innovation involved in this research to be widespread and for its positive results to act as an incentive to Saudi testing policymakers to investigate the feasibility of using the project package, or any of its built-in components, so as to improve HCD teacher assessment on a wider scale in all Saudi schools and in all other subjects. Moreover, training courses given by the MoE could be more focused on understanding item analysis parameter explanations and indications, rather than focusing so heavily on teaching how to calculate them correctly by hand.

1.10 Context

1.10.1 KSA

The Kingdom of Saudi Arabia (KSA) is a Middle Eastern country located on the Arabian Peninsula, as Figure 1.6 shows. The location of Al-Ahsa, the city where this research was conducted, is towards the eastern part of KSA (it is also known as Al Hufuf, as shown on the map). The 2004 Census recorded KSA's population as 22.7 million (growing at around 3% per year). The only language is Arabic and no second language is used. The country is less than one hundred years old, following unification (EIA, 2001) on 23 September 1932.

Figure 1.6: Map of the Kingdom of Saudi Arabia (KSA) (Gulf Law, 2011)

1.10.2 Education System

The education system in the KSA has four key stages: kindergarten (3 years for ages 3 to 5), primary (6 years for ages 6 to 12), intermediate (3 years for ages 13 to 15), and secondary (3 years for ages 16 to 18). The last stage is sometimes referred to as 'high school' in literature on Saudi education; thus the two terms, 'secondary' and 'high school', can be used interchangeably. Education is delivered in single-sex schools. Before 24 March 2002 (*Riyadh Daily Newspaper*, 2002, p. 1), there were two ministry-level organisations governing education: the Ministry of Education (MoE) governed boys' education from primary to secondary school, and the General Presidency of Girls' Education governed girls' education from kindergarten (in which both sexes mix) to secondary school. Following the royal decree of a merger on 24 March 2002, the Ministry of Education now governs education for both sexes. From the perspective of educational officials, the reason for this decision was to unify efforts by both organisations for similar goals, to lessen expenditure, and to unify educational policies (*Riyadh Daily Newspaper*, 25 March 2002). From many people's viewpoints, an 'undeclared' reason was to implement the experience that had been gained by the Ministry of Education in developing education and implementing new trends in educational issues that the General Presidency of Girls' Education had not accomplished (*Riyadh Daily Newspaper*, 26 March 2002). As a consequence, the transfer was to the former, the latter was cancelled, and both became the Ministry of Education (MoE). In addition, there are another two educational organisations, Higher Education and Vocational Education, which are governed by two different separate ministries.

1.10.3 Population Figures for Education in KSA

According to statistics for the academic year 2004/2005, the MoE in KSA provided education to 2.36 million male pupils through more than 197,000 teachers in 14,246 boys' schools, as well as 2.29 million female pupils through more than 228,000 teachers in 16,424 girls' schools. This yields a total of 4.65 million pupils taught by 425,000 teachers in more than 30,000 schools (MoE website). Education in KSA is free and provided mainly through state schools (referred to as public schools in KSA); therefore, the above figures indicate the huge responsibility of the MoE and the

challenge it faces every year to overcome obstacles and difficulties in running education over this wide and multi-provincial country.

1.10.4 Teacher Employment

Before the 1990s, due to a lack of Saudi professional teachers, most teachers came from other Arab states such as Egypt, Syria, Sudan, Iraq and Jordan. This deficiency in the number of Saudi teachers resulted in the MoE employing Saudi teachers immediately after their graduation from university, regardless of their level of competency in their field of specialisation, or their educational qualifications. As a result, many teachers nowadays are under-qualified, which has led to many undesirable practices. After a recent expansion in higher education, the new generation of teachers from the 1990s onward is now showing better educational skills. However, the effect of the inherited concepts and practices from the former teachers still play a major role in discouraging these new teachers from applying what they have learnt.

Another factor that is thought to be a barrier to educational improvement is that the incentive system for government employees (vis-à-vis teachers) does not associate yearly salary increment with employee performance, because it is based on automatic yearly salary increases. According to Friedman, Brinlee and Hayes (1980, p. 234), the "reward for professional growth should outweigh the reward for complacency." The 'automatic' salary increment might therefore have led to a decline in motivation to improve.

1.10.5 Assessment System

The academic year is based on two semesters, and the marking system is based on 100 marks for each subject, distributed as follows:

First Semester (50 marks):

15 marks for a mid-semester exam

30 marks for an end of semester exam

5 marks for a teacher appraisal

Second Semester (50 marks):

15 marks for a mid- semester exam

30 marks for an end of semester exam

5 marks for a teacher appraisal

The five marks are given by the teacher according to his or her "observation" of the pupil's activities throughout the semester. Indicators might include "homework assignments, classroom activities, their contributions during the lesson" (Student Assessment Document, 2000, p. 38).

There are no standardised tests in the Saudi setting, and the transfer from one year group to another is based on examinations set by teachers. There is only one national test, which is taken at the end of the third year of secondary school (year group 12, aged 18) and administered at the end of the second semester; thus, the final score is the sum of the teacher's test score from the first semester and the national test from the second semester. This means that the national test is relatively low in importance as far as the final result is concerned[16].

At school level, teacher-made tests are used for marking purposes only. There are no diagnostic analyses either in terms of pupils' learning or the qualitative characteristics of these tests. However, some educational supervisors do carry out some content analysis for teacher-made tests and summarise their findings in a report, which is then circulated to the teachers. Sadly, doing this depends on the individual supervisor and is not a routine course of his/her work. According to my experience as a former educational supervisor – and to some educational supervisors' reports that I have looked at for this purpose – I found noticeable shortcomings in teachers' tests. Among these are the intensive use of objective-type questions, such as 'true or false', 'multiple-choice', and 'comparison lists,' the limited use of HCD questions, and some errors either in the questions or their answers. A previous Saudi study found similar findings in this respect (Ayedh, 1993). From my perspective, I believe that the problem of excluding HCD questions deserves to be studied further as a priority.

1.11 Summarising Highlight

The proposition of this research is that higher level cognitive demand can be promoted by using IAT to make teachers more aware of what they are teaching and testing, and make it easier to undertake diagnostic assessments that promote higher cognitive level learning outcomes. Calculating IAT parameters is a difficult task which

discourages teachers from IAT as a diagnostic tool to improve learning. Thus the CAIAT software can help to overcome this difficulty and hence to motivate teachers to self-learn the skills required for improving HCD testing and instruction. If it is possible to increase the capabilities of teachers in writing HCD test items, and this leads to better understanding of how to teach and assess HCD learning objectives, this should lead to improvements in educational achievement. The MoE values this goal and thus the research is directly relevant to practice especially that no work of a similar kind has been undertaken in the region.

Chapter 2
Learning and Science Education

Why learning? As Jon Scaife puts it, “The more we know about learning, the better we can teach” (Wellington, 2000, p. 64). Learning should be the final aim of any educational system, because education is intended to assist individuals to make themselves ready for interacting with society – not just following its rules and traditions, but rather improving, developing, exploring, and inventing. This cannot be achieved without the construction of critical and thinking minds or building up confidence to change and innovate, for which learning is the major effective vehicle. In order to identify the relationship between learning and science education, I need first to highlight the main theories that explain learning as a concept and a cognitive process. I will then present several issues regarding science education, focusing on the relationship between teaching science and assessment and on change attempts within science teaching. I will then elaborate on one distinguished project, known as CASE, which promotes learning in the science curriculum by triggering thinking on HCD levels. This project was initiated in the UK, but has been adopted in some other parts of the world. In the present research I will make use of CASE partially, whereby its conceptual ideas will be employed within the training course that represents an important part of this research’s treatments.

2.1 Learning Theories

Bloom taxonomy is a pivot for this research's theme. The HCD levels of the taxonomy require teaching strategies that enable the learner to achieve the thinking skills that these levels require which requires a state of art of teaching. Efforts of educational and psychological theorists in this regard aim to provide us with an insight about learning process of the individual behaviouraly, mentally, psychologically or socially wise. I am going to illustrate learning theories chronologically, classifying all into three popular categories: behaviourism, cognitivism, and constructivism . This is to illustrate how this area has developed to conclude by a vision that outlines how could these learning theories be employed for effective teaching design that by which teachers could guide pupils to accomplish skills of the HCD levels.

2.1.1 Behaviourism

The history of Behaviourism is closely linked to the history of educational psychology. It is thought that Aristotle is the pioneer with his essay "Memory." Following philosophers such as Hobbs (1650), Hume (1740), Brown (1820), Bain (1855), and Ebbinghause (1885) continued his teachings. Later key 'experimental' behaviourists are Pavlov (1849-1936), Watson (1878-1958), Thorndike (1874-1949), and Skinner (1904). The main concept is based on stimulation and response, where the mind functions as a respondent to suitable stimulus. Those responses could be measured quantitatively - hence the experimental aspect of this approach. A well-known experiment is Pavlov's classical conditioning in which he rings a bell for a dog just before serving its food, resulting in salivation whenever it hears the ring again. No attention is given to mind processes or effect of environment. Thorndike's popular saying, "Anything that exists, exists in a certain quantity and can be measured" (Custer, 1996), reflects his philosophy towards quantification of human behaviours . His theory, *Connectionism*, implies that connection between stimulus and response results in learning. Watson presented the link between stimuli and response in terms of emotional effects. Skinner's behavioural shaping is the most mature view of the conditioning approach. He articulated the conditioning concept in a practical technique to shape behaviour in a series of planned gradual uses of stimuli to shape the behaviour. (Child 1983: 94-102, Salih, 1988: 377-392 and Mergel 1998).

2.1.2 Cognitivism

Limitations of Behaviourism come from being unable to explain some situations that oppose its thesis. Examples are that children do not acquire all behaviours that have been reinforced, or adversely, they might show a new behaviour that they had observed weeks ago without any reinforcement (Dembo, 1994. As sited in Mergel, 1998). Cognitivism arose in 1920s by Jean Piaget to cover issues like these. His first article was on the psychology of intelligence (Boeree, 2006), followed by the theory of cognitive development which is characterised of the concept of "schemata" which are schemes of how the individual perceives the world in terms of four developmental stages: sensorimotor stage (years 0-2), preoperational stage (years 2-7), concrete operational stage (years 7-11), and formal operational stage (years 11-adulthood). Implications about memory: how it works, how it could be employed for better learning,

what affects its function, and similar issues have coloured Piaget's vision of learning (Abu Hatab and Sadeq 1992: 266-283).

His work is considered one of the most influential contributions to developmental psychology. In Education however, the spectacular aspect of Piaget's theory is the "child-centred" approach. It is believed that his ideas have had a considerable impact in Britain's education (Wellington, 2000: 65). Piaget's four stages of cognition are illustrated next.

1. Sensorimotor stage (years 0-2): Infants use senses and motor abilities to understand the world (Boeree, 2006).

2. Preoperational stage (years 2-7): In this stage, the child: will be able to make use of symbols' concept, understand scale of time in terms of past and future, and is egocentric (Boeree, 2006).

3. Concrete operations stage (years 7-11): More developed, logical, and mental operations such as conservation, classification, and seriation[17] are seen in this stage. (Boeree, 2006).

4. Formal operations stage (years 11-adulthood): Some recent researchers such as Shayer et al. (1976) and Lawson and Renner (1978) indicated that most children reach formal thinking later than 12 years old; and moreover some do not reach it at all (Wellington, 2000: 66). In general, we should denote that this level of thinking is characterised by its difficulty, even for mature people, as Boeree highlighted:

> Even adults are often taken-aback when we present them with something hypothetical: "If Edith has a lighter complexion than Susan, and Edith is darker than Lily, who is the darkest?" Most people need a moment or two (Boeree, 2006).

This level includes the use of logical operations in abstract rather than concrete situations. This is called "hypothetical thinking" which is organised systematically and its main purpose is explanation, and is characterised by using 'variables' and mathematical concepts such as ratio and proportion (Wellington, 2000: 66). Given that these are obvious aspects in science education, the call for teaching techniques for formal thinking in science instruction is a must.

2.1.3 Constructivism

Lev Vygotsky (1896-1934) is the most constructivist figure who considers learning as a process that is unconfined when delivered in a responsive social context. When children are given enough assistance, “scaffolding learning,” then they can show more competent performance (NASP, 1997: 2). The Constructivist theory includes a relationship between what is being taught and pupils' experiences and personal purposes; therefore, “students are builders of knowledge who actively construct the meaning of their lessons on the foundation of both their past experience and their personal purposes” (Henderson, 1992: 5). In this setting, this type of learning tackles HCD levels. When teacher-made tests appreciate this approach and give enough attention to HCD test items, then pupils are expected to respond actively to the requirements of such interactive learner-based instructional process. There are three assumptions underlying the adoption of constructivist pedagogical approach as suggested by Confrey (1990):

> 1. Teachers must build models of student's understanding … The result will be that a teacher creates a "case study" of each student.
> 2. Instruction is inherently interactive … Teachers must be prepared to revise their own beliefs or to negotiate with the student to find a mutually acceptable alternative.
> 3. Ultimately, the student must decide on the adequacy of his/her construction (As cited in Brownstein, 1997: 12).

Wang, Haertel, and Walberg (1993) carried out a meta-analysis which found that the significant factor which affects learning is teacher-student interactions. Schon (1989) indicates that reflection-in-action (RIA) represented by interactive assessments is the important component to student learning (Brownstein, 1997: 3).

Zone of Proximal Development (ZPD) is one of the common resulting ideas of Vygotsky's constructivist theory of learning. This zone is the gap between what the learner can do by her/his own for a definite learning task and what she/he can do with the assistance of an instructor. Within ZPD, teachers assess students by determining what kind of help they need to complete a task successfully (Brownstein, 1997: 4); therefore, assessment represents the spine of the constructivist pedagogical approach. HCD questions come on top of this assessment as it tackles students' thinking in the first place and facilitates the required social interaction needed by the module. *Scaffolding* is parallel to ZPD concept which Wood, Bruner, and Ross introduced in 1976 to describe

an adult providing children with aids for learning how to do things they were not able to do alone (Hobsbaum, A., Peters, S. and Sylva, K., 1996 in Kristinsdóttir, 2001).

As a sociocultural approach, it is thought that Vygotsky's theory has incorporated elements of Marxism, a philosophy that emphasised *socialism* and *collectivism* contrary to *individualism* in the sense that the success of any individual was seen as reflecting the success of the culture. Intellectual abilities are seen by Vygotsky as being rearing culture-specific (Vasta, R., Haith, M.M., Miller, S.A., 1995 in Kristinsdóttir, 2001). *Dialectical process*, where the child learns through shared problem-solving experiences with someone else, is another pillar of Vygotsky's vision with a stress on language dialogue as a major median of this dialect. (Kristinsdóttir, 2001). Sexton (1997) labels the present era as postmodern/constructivist which pay special attention to epistemological issues in the sense of *how* people know rather than just *what* they know. Furthermore, the nature of meaning is relative; phenomena are context-based; and the process of knowledge and understanding is social, inductive, hermeneutical, and qualitative (Raskin, 2002).

Given that constructivism is a recent trend in learning philosophy, many sub visions have arisen. Among those are: *radical constructivism* in which Von Glaserfeld emphasises that individuals use the understandings they create for their life, regardless of its congruency with an external reality; *social constructivism* which emphasises the "primacy of relational, conversational, social practices as the source of individual psychic life" (Stam, 1998: 199, as cited in Raskin, 2002); and *personal constructivism* in which "people organise their experiences by developing bipolar dimensions of meaning, or *personal constructs* and they then test and revise these constructs" (Raskin, 2002).

2.1.4 Discussion and Conclusion of Learning Theories

Stage theory of Piaget has been questioned by number of writers as Driver (1983) highlighted:

> Currently, there is some controversy as to the validity and utility of the so-called stage theory. It is recognised that the ability of a pupil to use certain logical operation, for example proportional thinking, depends on his familiarity with the context within which a task is set. Pupils may control variables competently in one task but not in another. This means that it is pupils' behaviours and responses which can be labelled as fitting a specific stage, not necessarily the pupils themselves (Driver, 1983: 56).

However, this type of questioning is seen to be directed at interpretations of Piaget theory rather than its original concept of stages. These interpretations lie in three positions: the *structuralist*, in which, learning is age-dependent, the *memory-based* in which learning is age-dependent but limited by the capacity of working memory, and the *sequentialist* which proposes that learning depends primarily on previous knowledge of the individual (Driver, 1983: 57-59). Similarly, constructivism and the learner-centred approaches have been attacked. Ledda (2005) reviewed three books written by teachers and showing their refusal to these concepts and calling for acknowledging the need to knowledge transfer concept. Inevitably, views about a philosophical and new trend such as constructivism are very likely to last especially when we consider the evolving revolution of information age:

> The socio-cultural aspects in Vygotsky´s theories are interesting when analysing the learner in the information age society. How do we educate the child raised in a world of instant information, where interactive technologies have led them to believe they can act on the world with the press of a button? (Kristinsdóttir, 2001).

I think that such authors have understood that the new trend entails neglecting wholly every shape of former trends, i.e. behaviourism and cognitivism. Actually, each vision could function for a purpose or a degree of purpose. For example, behaviourism led to the use of instructional behavioural objectives, which is still valid as one important instructional instrument. Also, implications of rewards and reinforcement are still considered within the constructivist instructional setting. Conciliation of this entire heritage is the proper and wise treatment of the multiplicity interpretations of learning processes. This vision entails adopting a major trend of the instructional activity (such as constructivism) with the appreciation of congruent concepts of other approaches. The approach of the present project training package subscribes to this vision where it adopts collaborative learning as a means for crossing the zone of proximal development that the constructivist approach implies. It also encourages teaching thinking and considering contextual variables and situational needs which are informed by behaviourism and cognitivism.

2.2 Learning and Teaching Styles in the KSA

Al-Nassar & Al-Sughayyer (2002) surveyed 350 out of 16,768 Saudi teachers at all key stages of Riyadh schools to determine what instructional practices they applied

according to three learning theories approaches: behaviourism, cognitivism, and humanitisim. They found on average that, generally, teachers' practices complied with learning theories, but the authors commented that this finding do not necessarily reflect the level of the teachers' attainment in relation to the concepts or theoretical knowledge underpinned by those theories. The study showed that humanitisim came through strongest, followed by cognitivism and then behaviourism. In terms of differences between key stages, there were no significant differences. For level of qualification, in terms of cognitivism, those who graduated from university demonstrated better levels than those less qualified, while in terms of the other two approaches there were no significant differences. The study recommended training teachers in this area and providing them with related training materials.

Al-Mani' (2005) explored which teaching and learning style preferences are applied in Saudi middle schools. She found that the learning styles most preferred by students were verbal interaction, learning by doing, and collaborative learning. Conversely, learning by rote, autonomous learning, doing exercises, and solving problems were the least preferred by students. The preferred learning styles included some motivating dimensions such as interest, challenge, choice, and enjoyment. However, these were found to be rarely used in the most common teaching methods, which indicates that teaching styles were not meeting students' learning style preferences.

Consequently, Al-Mani' recommended that instructional strategies congruent with students' preferences should be implemented. Since motivating dimensions are core aspects of humanitisim, the latter finding might sound contradictory to the preceding study of Al-Nassar & Al-Sughayyer (2002), which indicated that humanitisim was top amongst practiced approaches by Saudi teachers. Nevertheless, this could be interpreted by the fact that the differences between the three means in Al-Nassar & Al-Sughayyer's (2002) study did not split the three theoretical approaches apart too much; their values were 4.15, 4.02, and 3.96. According to Badr (2006), teaching maths in Saudi School depends on classical instructional methods, while problem-solving and discovery methods are used moderately.

Despite these studies indicating that teachers' practices and pupils' learning styles did not satisfactorily meet contemporary trends in pedagogy, many other local studies piloted the use of modern methods of teaching, such as constructivism, exploration, problem-solving, etc. The aim was to examine the impact of the extent to

which these modern methods could succeed in a Saudi school and revealed highly encouraging findings. To mention a few: Al-Awadh (2007), Badr (2007), Al-Aklubi (2008), and Brikeet (2009). To conclude, teaching styles in KSA are mostly typical/classical methods and not meeting students' learning needs; however, the reported attempts towards utilising the modern instructional styles are encouraging. I believe that this area is not researched enough thus needs to be given a better interest and focus.

2.3 Learning in Science Education

2.3.1 General Overview

It is worrying that science education might be left behind as a result of the big gap between the speed of science and technology advancement and the slow development of science education. Holbrook (2003) considers that science education's research lies within two domains: interest in HCD and teaching context. Therefore, 'Education Through Science' means preparing students for everyday life within society, which could establish skills for solving problems and making decisions, and thus improve the quality of life. This requires paradigm changes in science education represented by relevance to one's life, avoiding memorisation as an instructional essential, and constructivism as a heart for learning (Holbrook, 2003). Furthermore, current science education practices is said to be characterised by the classical transfer of frames of knowledge.

> The grounds for accepting the models proposed by the scientist is often no different from the young African villager's grounds for accepting the models propounded by one of his elders.....For all the apparent up-to-dateness of the content of his world-view, the modern Western layman is rarely more 'open' or scientific in his outlook than is the traditional African villager…Science is one of the last intellectually dogmatic and authoritarian disciplines on the school curriculum (Osborne, 2005: 7).

Osborne highlighted 5 Aspects of good practice in teaching science that are illustrated in Table 2.1 below. Among those, she stressed the importance of teachers having a proper conception of the nature of science and "what it means to teach science" (Osborne, 2005: 27). She also outlined three functions for science education: "to show that science offers a means of creating new knowledge, to educate the next generation of scientists, and to educate the future citizen" (Osborne, 2005: 1)

Table 2.1: 5 Aspects of good practice in teaching Science (Osborne, 2005: 18)

Aspect	Traditional	Learning-based
Teachers Knowledge and Understanding of the Nature of Science	Teacher is anxious about their understanding of nature of Science	Confident that they have a sufficient understanding of nature of Science
Teacher's Conceptions of Their Own Role	Dispenser of knowledge	Facilitator of learning
Teachers' Use of Discourse	Closed and authoritative	Open and dialogic
Teachers' Conception of Learning goals	Limited to knowledge gains	Includes the development of reasoning skills
The Nature of Classroom Activities	Student activities are Contrived & inauthentic	Activities are owned by students and authentic.

The use of effective assessment materials is thought to lead to an effective inquiry-based science instruction. Research reported that participating physics students in a reflective assessment activity had higher results than those of control classes (White and Frederiksen, 1998). Teaching thinking's conceptual framework consists of different principles or ideas such as HCD, metacognition, reasoning skills, and critical thinking. Alfred North Whitehead (1929) highlighted the core goal for teaching thinking: "what you have learned would be useless unless you get your books lost, burn your memos and forget all of what you have memorised for testing purpose," which indicates the great role of thinking as a final aim for learning, as it is not what we 'know' but rather it is what we 'can know.' Although solving problems is one of the most prominent aspects of teaching thinking, finding problems on the other hand is considered a more powerful instrument for stimulating creativity and developing thinking practices. Arlin (1990) in his opinion for understanding wisdom suggests that problem finding is more significant than problem solving (Brooks, 2002). I believe that although it is difficult to find a substantial evidence for this view, one should consider both ways essential.

2.3.2 Science Education and Change

2.3.2.1 Overview

Beate Davies (2004) studied the British science curriculum in terms of teachers' perspective about how they teach compared to how they are required to teach. He interviewed a small number of science teachers with a wide variety of experiences, and implemented concepts of *intended*, *implemented,* and *attained* curriculum. He

considered assessment influence on curriculum as a consequence of targeting accountability, which in turn affects pedagogy (Davies, 2004). He found that there are constrains coming from a lack of time and a lack of educational authorities' trust in teachers' abilities to make decisions about their teaching indicating a decline in their professional status. This has been analysed as a consequence of the content-laden curriculum of the national curriculum in the UK, pressure to provide measurable value-added components, administrative burden, and the need to maintain the school's league-table position. Practical work was "sacrificed" with a focus on the traditional style of one-way instruction. The point of interest here is that he concluded that these teachers were teaching to the test, and hence called for further studies at the national level in the UK, and to give teachers more trust and empower them for better constructivist teaching (Davies, 2004: 64-70). From my perspective, I also think that the size of science curriculum content should be looked at since this study revealed its role. However, Davies cited a section of Alan Bennett's play *The History Boys* (2004) in which Irwin is teaching for the examination and Hector is teaching 'for life.' I find this citation as interesting:

> '**Irwin:** Education isn't something for when they're old and grey and setting by the fire. It's for now. The exam is next month.
> **Hector:** And what happens after the exam? Life goes on.' (p49)
>
> Later in the words of the headmaster, Bennett characterises the problem, when accountability and assessment lead pedagogy instead of the *intended* curriculum:
>
> 'Shall I tell you what is wrong with Hector as a teacher?
> It isn't that he doesn't produce results. He does. But they are unpredictable and unquantifiable and in the current educational climate that is no use. He may well be doing his job, but there is no method that I know of what enables me to assess the job he is doing. There is an aspiration, certainly, but how do I quantify that. And he has no notion of boundaries. A few weeks ago I caught him teaching French. French!' (p67) (Davies, 2004: 70-71)

The use of technology in school is one of change practices that is seen predominant recently due to the wide spread and express development in educational technology's application especially those related to computers. Science education has received part of such change attempts. The study of Gujski and Ben-Peretz (2005) tackles physics teachers' PD in which they introduced physics teachers to PD program for integrating computers in physics teaching. Their hypothesis was: the more the teacher experienced the lower his/her professional uncertainty, which was also

accompanied by a consequent decrease in uncomfortable feelings. They demonstrated a number of visions towards uncertainty, but the common thread was lack of knowledge or confidence in the possessed knowledge. In 1993, a group was formed in a workshop for the integration of computers in physics teaching in the laboratory. From then to 2000, less than 10 teachers out of 20 remained as participants; and out of these, the researchers used "group case study" to interview seven physics teachers and the main instructor who coordinated the training program. Moreover, they analysed documents, previous interviews, and evaluation reports from the first years of the training; then they used "theory driven thematic analysis" for analysing respondents' responses, which revealed that the teachers did not feel uncomfortable in uncertain situations. They demonstrated that the teachers confidently reached a high level of professional knowledge and most of them became expert teachers because they used those uncertain situations to improve their teaching. They attributed this finding to what is reported by Gabella's (1995) study; in which, secondary school science teachers, instead of perceiving uncertainty as distracting, they considered it challenging which elicited better learning. Nevertheless, they commented that for teachers who did not participate in training courses, particularly elementary school teachers, there might be different situation, such as that found by Lange and Burroughs-Lange (1994), in which, professional uncertainty decreases with experience and hence they called for further relevant research (Gujski and Ben-Peretz, 2005).

2.3.2.2 Cognitive Acceleration through Science Education (CASE)

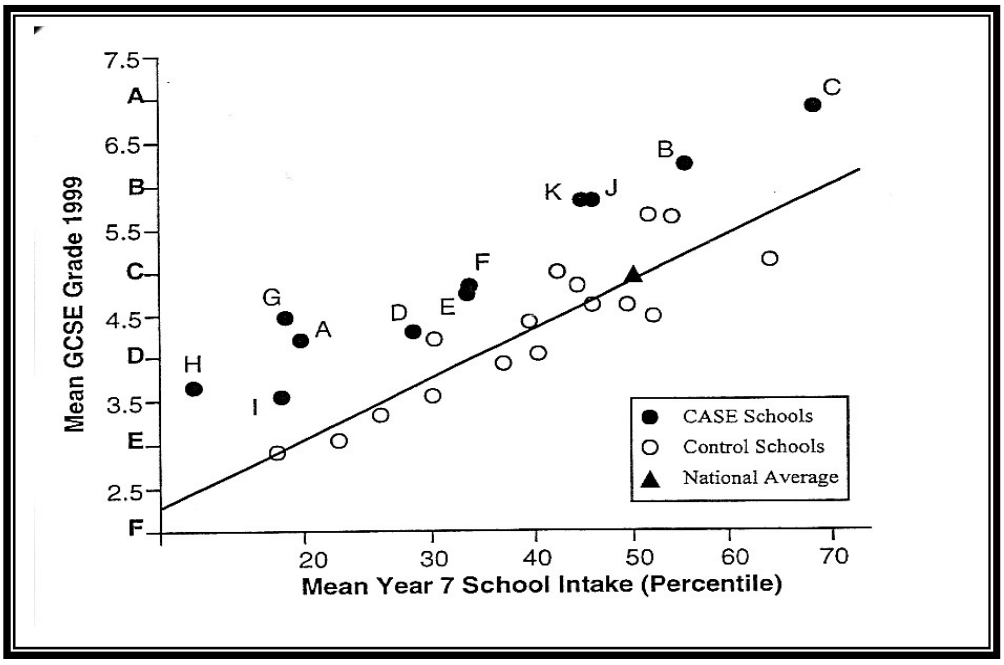
Cognitive Acceleration through Science Education (CASE) is one of the distinguished projects for eliciting learning in science education and provoking thinking. It is an ambitious project that was established by Michael Shayer from King's College, London School of Education in 1981 as a principle, and then with Philip Adey and Caroline Yates in 1984-87 as a practice (Adey, 1995: 1). Findings of the project showed that cognitive acceleration (CA) approach has a positive impact on pupils' learning, which has been measured by the added value of the intervention to the project schools' GCSE results compared to those of the control-group schools. The intervention includes special design for lessons and training for teachers on skills that evoke learning better and help pupils to develop their thinking skills. The theoretical framework of CASE is based mainly on the works of Piaget and Vygotsky. One good and short explanation of the idea is that of Adamczyk, the representative of CASE project at

Sussex University, who said that they employed the opposite of the usual setting of teaching -- teachers introduce learning skills using content, instead of the usual setting of introducing content by utilising learning skills. That is to say that the subject matter of science in CASE are just examples of what learning skills the teachers want to train or teach pupils to do or to acquire; in short: I describe the process by the form of 'upside down pyramid.'

2.3.2.2.1 Impact of CASE

The substantial and long-term effect of CASE method on pupils' achievement can be seen from Diagram 2.1 representing results of 11 experimental schools (A to K black circles). White circles are the mean percentiles of schools of the control group. These are for Year 7 intakes, plotted against the mean of their GCSE and distributed around the regression line that represents the national average. Black circles as located above the regression line show that all of CASE schools have gained better GCSE results after 3 years of application (Shayer, 2000: 2-5).

Diagram 2.1: GCSE 1999 science mean grades: added value (Shayer, 2000: 4)



Moreover, teachers who applied its lessons illustrated some short-term benefits of CASE. Among those is a Monifieth Secondary school's teacher (School Material, 2001: 3):

> The pupil benefits are that: they are more confident in attempting to answer questions, they appear to enjoy these lessons, they found writing up investigations was easier to complete, they are more willing to make predictions … and to try to explain … The positive outcomes for me are that: it made me think about the questions asked … and it has had the major advantage of giving me back some of my lost enthusiasm for teaching.

Also, he commented that "one of the hardest areas found at the beginning was asking the correct questions for metacognition and bridging," which points to the major role of classroom assessment. This highlights the priority and importance of training teachers on how to construct such HCD questions.

2.3.2.2.2 CASE Lessons

CASE lessons focus on themes of thinking skills which are called reasoning patterns (Adey, Shayer, and Yates, 2001: 8-10). In CASE lessons, the Piagetian approach is embedded within the activities, while Vigotsky's approach is seen in the social interactive setting of instruction. There following five pillars for teaching by the CASE method form phases of the teaching process in CASE (School Material, 2001: 2):

1. Concrete Preparation: where the teacher demonstrates the problem, terminology to be used, apparatus, and shares with pupils planning for the lesson activities (MSIS, n.d.: 2).
2. Cognitive Conflict: When a new concept is introduced to pupils contradicts a former one, then this conflict is supposed to elicit thinking either for resolving it or developing new ways of thinking.
3. Construction or reconstruction: occurs when pupils reach to the final picture where "construction of knowledge is often a social process, happening between teacher and pupil and pupil and pupil" (MSIS, n.d.: 5).
4. Metacognition: is a reflection to Vygotsky's ZPD theory. In this phase, there is emphasis on thinking about thinking alongside giving enough time (15 seconds) for the pupil to analyse and answer (School Material, 2001: 2)
5. Bridging: bridging back to a former content that the current activity builds on, and bridging forward to a new area of science or other context as well.

2.3.2.2.3 Discussion and Conclusion about CASE

There are some people who are sceptical about any claims of long-term effects as CASE project claims (Wellington 2000: 96). However, the increasing number of UK schools that have adopted the CASE's Thinking Science package alongside its science

courses points in favour of CASE's merit. Reinder Duit thinks that the pedagogical approach of cognitive conflict is failing. His opinion comes from the fact that there should be some pupils whose personal conceptions are too strong to be shaken by a conflicting situation or experiment (Jaworski, 1995). Although this could be true for some pupils, I believe that this does not necessarily denote an end to this pedagogy, as much as a matter of concern that should be taken into account when adopting it with some further 'remedial' action that should be applied for lessening this effect with this little number of pupils.

The setting of CASE lessons with its applied vision to collaborative learning starting from *concrete preparation* and ending by *bridging* have been employed in the present project through the training course that was made to prepare the teachers to teach better on HCD level. The training focuses on a number of procedural arrangements during teaching such as group work, reinforcement, feedback, and teacher discussions with pupils. I have to highlight that CASE's essential aim is beyond the present project's limits and focus, but, I have made use of its inclusions by various approaches alongside the training which I have indicated to. I have made use of the optimal questioning practice indications of CASE and its inclusions regarding the social arrangement of group work. I also presented CASE intervention to the sample teachers in order to gain their confidence in the merit of the new introduced 'teaching thinking' approach. In this respect, some DVDs that explain CASE were translated into a written Arabic script on the screen and played to them, which not only benefited the trainees, but also simplified for the trainers a way they could present these abstract skills to the sample teachers.

Chapter 3
Educational Change

3.1 Overview

In recent years, the literature on business administration as well as education has become rich in theories, studies, and academic papers regarding change. This invites me not to elaborate on introducing basic concepts as much as presenting points that underpin the present researched attempt to change and highlighting those areas that could aid to explain findings and inclusions of this attempt. This presentation will progress from general to specific, where I will begin with change theories in general then move to organisational change followed by theories of personal change focusing on resistance to change and motivation. I will then elaborate on professional development (PD) focusing on action research (AR) since it represents a pivot in this regard. I conclude by a presentation of Diffusion of Innovation theory, DoI, that explains how technological innovations diffuse since this study has a technical side.

Change is characterised by being system disturbing, nonlinear, personal, organisational, assuming resistance and conflict, and multidimensional. The principle feature of change is that people factor is the major role player. The breakthrough in this term is that involvement of individuals inevitably creates uncertainty within change process. Many writers think that people involved affects its progress in the first place rather than the content of change. Actually and in terms of Morrison rhetorical expression: "change changes people but people change change" (Morrison, 1998: 14-15). Nevertheless, it is important to look at management factor, considering it is a ground for this presentation.

Eraut (2004a: 112) describes the research literature of change being like a "smorgisborg"[18] of theoretical abstracts that one chooses from regardless of their context or theoretical underpinnings, and thinks that many theories are complementary rather than oppositional. He recommends House's (1979) three dimensions that characterise all change processes, work at all levels, and conform with what Bennis, Benne & Chine (1961) suggested in this term; these three dimensions are: technological, political, and cultural. However, Eraut adds two more emotional dimensions:

> Two contrasting, but related, adjectives that come to mind are 'comfortable' and 'confident'. 'Comfortable' has overtones of maintaining one's current spread of activities and relationships, whereas 'confident' suggests a willingness to take on new challenges. Both relate to the emotional dimension of change, which appears to be missing from the three paradigms of Bennis et al. (Eraut 2004a: 114)

In terms of "comfortable," he commented that Hertzberg's (1966) classical studies highlighted incompetent management or bad working conditions as barriers to motivation but not an end to this respect, where making these factors available would not necessarily motivate individuals. Challenge is also considered the strongest motivating factor. Eraut's research on mid-career learning confirmed this and noted that:

- Successful completion of challenging work was a major contributor to confidence in one's own capability; and
- such challenges were less likely to be accepted, or even noticed, when there was little support in the form of encouragement and constructive feedback (Eraut 2004a: 115).

Moreover learning is found to be affected by situational factors such as climate and culture of the workplace. Formal learning in settings away from workplace is required for new practices of continuous PD but will also depend on following informal learning in the workplace (Eraut 2004a: 112). Change in education is a continuous and viable process since education is a reciprocal interaction with the changing social system in which it functions. Many writers have suggested different models for managed change processes that mostly meet at some point or another and generally attempt to systemise the process either for implementation or analysis. I am going to illustrate some of these models shortly.

3.2 Perspectives of Change

3.2.1 Organisational Change

Hersey and Blanchard (1988) suggest two strategies as shown in Figure 3.1 below. The first is the 'participative change cycle' motivated by the *personal power* and is a bottom-up cycle. The second is the 'directive change cycle' applied by the *position power* and is a top-down cycle. Personal power starts from knowledge that stresses a high role of learning and professional development. In this study, I am drawing on this form of planned change to gain its rich and reliable effect.

Figure 3.1: Two strategies of change (Adapted from Hersey and Blanchard, 1988: 340-341).

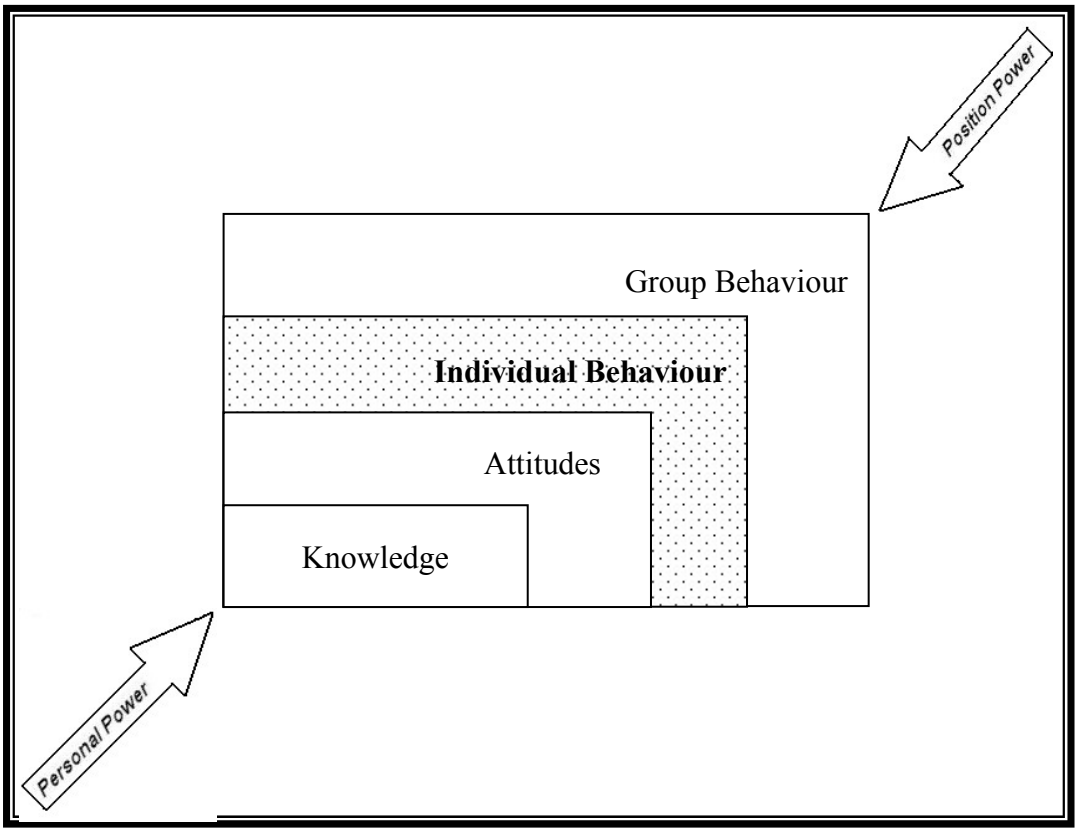


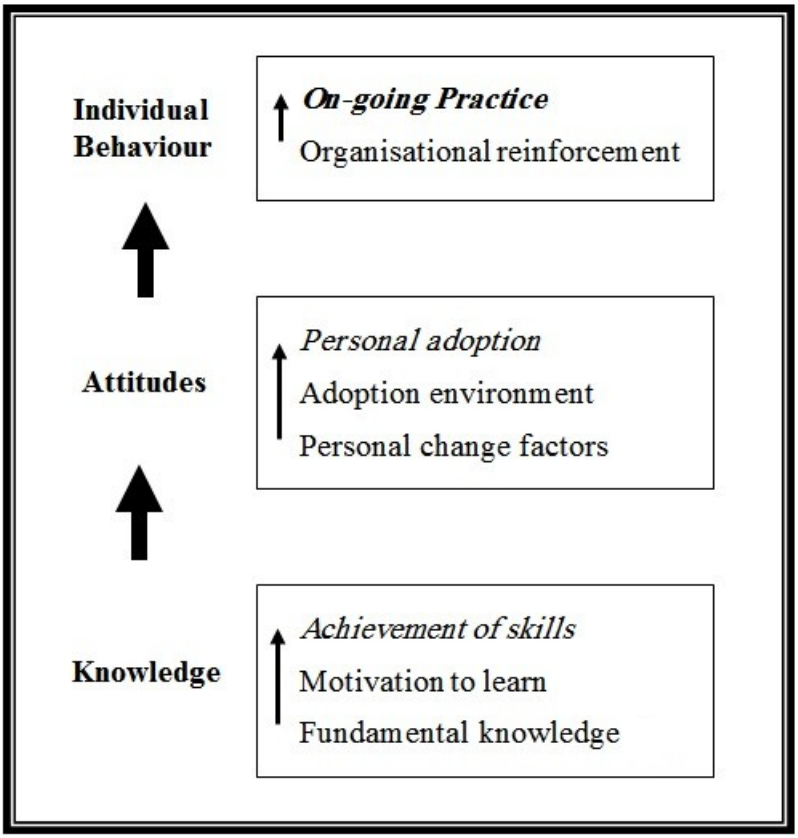
Figure 3.2 provides my model that elaborates how personal power flow through the two areas of knowledge and attitudes in order to establish the new practice on the level of individual behaviour. It expands knowledge into three gradual steps. The first is the fundamental knowledge (FK), in which concepts that underlie the new intervention need to be introduced to individuals to get their initial attention. There should be some aspects or gains that motivate individuals to learn the new idea because without this motivation, the FK that is received will be no more than a cultural enjoyment. Finally, collective or individual learning yields the needed skills by which the individuals are able to perform as required.

Having skills accomplished by the individual does not necessarily mean that change will occur because individuals have to apply what they have learnt, which requires reaching a good level of adoption as the second stage of the model includes. The first part of the second stage, *personal change factors*, is articulated by personal change theories that describe factors that affect how the individual proceeds to the new change aspect. These factors should be available especially those coming from the

design of the new intervention and the setting of its implementation. The second part, *adoption environment*, requires that a good environment should exist to promote reliable and an on-going application; otherwise, any internal personal positive intentions towards the new idea will vanish over time. With the existence of both *personal change factors* and *adoption environment*, the intervention's amount of implementation has become rich enough to inculcate a personal positive attitude towards the new intervention and mostly become part of the individual's value system which represents an optimal personal adoption state.

Having reached this point does not also guarantee a continual practice unless there is a supportive or reinforcing effort from the organisation which is the final stage of the model. This represents a bridge for institutionalisation through which a new idea will become an on-going practice. I will elaborate later how these three stages of the model, knowledge, attitude, and individual behaviour, will be reflected in the present researched change attempt.

Figure 3.2: Personal power model for establishing individual behaviour



It is essential to denote that my research topic selection focusing on personal power path to achieve personal change should not restrict change to the individual. It is rather to take the advantage of the optimal opportunity in the personal power path to establish change over the whole organisation, which is an ultimate goal that is represented by the 'Group Behaviour' as the model in Figure 3.1 illustrates. Nevertheless, the limits of this research setting do not permit to report the targeted change on the organisation level. In the following section, I will illustrate organisational change approaches in order to not tackle these as a framework for the design of this research. Instead, I will describe where my model fits within such a macroscopic perspective of change.

3.2.1.1 Approaches to Organisational Change

The *systems* approach is concerned with tackling change through determined steps that are carried out by systematically planned rules and procedures in which the operational functions of the organisation, or system, are the key issue. The *Phenomenological* approach, on the other end of the continuum, does not pay attention to this level of determination and tackles the issue of change from a very broad view. It looks at the organisation as coming from people who relate to it and realities govern their interaction; thus organisations do not exist according to their constructs, but according to what people do within them. Therefore, the top-bottom change path subscribes to the systems approach while the bottom-up change path subscribes to the phenomenological approach.

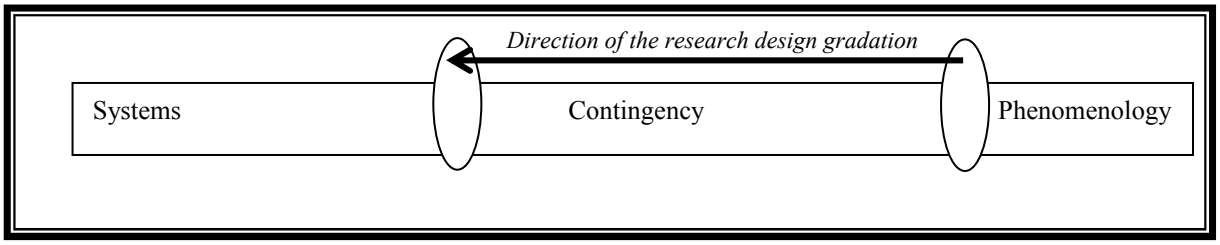
Systems approach cannot be overlooked by managers or management theorists, as variables, components, and aspects of their jobs are formed in systems (Koontz and O'Donnell, 1976: 14). At the same time, the philosophical essence of phenomenological approach may be found in Wittgenstein doctrine which implies that "the criteria of knowledge emerge out from multiplicity of practice in which we are immersed" (Hodges and Lachs, 2000 :9). *Contingency* approach lies in the middle, thus makes use of both trends' components; it aims to balance their benefits and problems.

In behavioural sciences, a contingency is a "relationship between behaviour and its consequences" (Berman, 1971: 4). Contingency theory in management planning is based logically on the following notion:

> A contingency approach attempts to join learning with action, to move incrementally toward effective and efficient implementation, based on knowledge and experience gained through interaction with participants and beneficiaries (Rondinelli 1983a; 1983b in Rondinelli et al. 1990: 18).

Contingency theories depend on incremental adjustments, which come from learning during implementation. They render the degree of the need of this adjustment to the level of uncertainty of the environment and complexity of innovation or change. (Rondinelli et al. 1990: 25). From this, the key issue of the theory is either to cope with uncertainty, or to cope with complexity of the innovation, yet both issues call for an adaptive strategy rather than mechanistic strategies such as that found in systems approach (Rondinelli et al.1990: 74). Due to this interposition, this is the approach which will be seen most influential by my design. However, the technical orientation of the discussion might sometimes give the impression that this is a systems approach. Also, it might be suggested that the early stages of the research project's strategy tend to favour the phenomenological approach and the late stages tend to favour the systems approach. To an extent this is true, since the project design is dynamically progressing from a mid-position between phenomenological and contingency approaches, towards another mid-position between contingency and systems approaches. Figure 3.3 illustrates this.

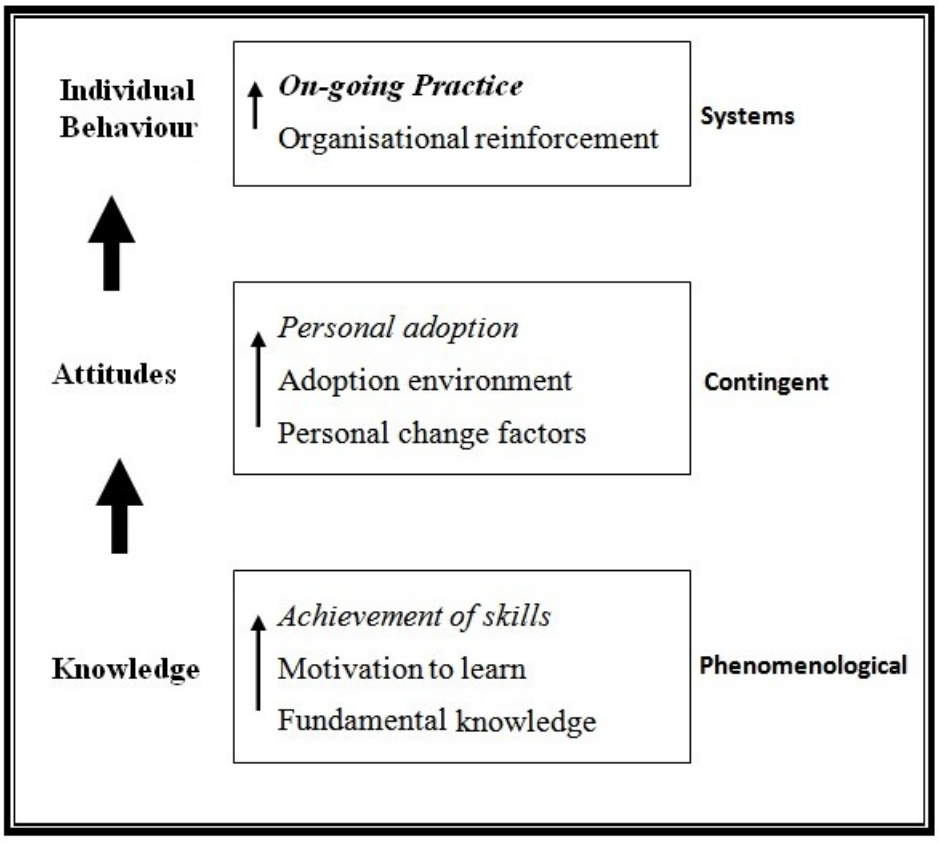
Figure 3.3: Illustration of the research project's design on the continuum of change approaches



When this vision is applied to the model of Hersey and Blanchard, it shows just how much knowledge does play a major role in disseminating the new concepts and skills of the HCD/CAIAT project. The outset is by initiating a sort of phenomenological aspect of the targeted change. As the scheme of the work plan will show, training will be a very helpful vehicle for this purpose. The sequence of stages of the project aims to create a teachers' personal positive attitude towards what they have learnt in training. It is at that point that their behaviour would be expected to be

coloured gradually by the new trend then they continue their PD efforts on a self-learn basis so that they improve their skills overtime. This stage is characterised by 'self-learning' which portrays the contingent approach; therefore, I have called this stage in this project's work-plan the contingent stage. This gradation, represents a gradual progression from the position near to the phenomenological approach towards the contingent position. For this project's plan however, it will be headed in the direction towards systems approach but it may not exceed the contingent stage too much because an institutionalisation is very likely the end result of this path when the MoE adopt the project for a sustainable implementation. Figure 3.4 is a re-draw of the *personal model* which I suggested for this project showing how the model fits with the vision of approaches to change as mentioned above. The *Knowledge* and *Attitude* stages with their identical stages *Phenomenological* and *Contingent*, are respectively the two levels of the model within which the project will function. The third level is achieved by the recommended official implementation to be undertaken by the MoE.

Figure 3.4: Personal power model with approaches to change



3.2.2 Personal Change

If we were to describe change as a personal process, then it is that it makes people think, believe or act differently and therefore be likely to feel different. Judson (1991) mentioned five types of reactions to change: acceptance, indifference, passive resistance, active resistance, and sabotage (Morrison, 1998: 122). Resistance could include vocal protests and attitudinal outcomes such as withdrawal and reduced commitment (Judson 1991; in Souchon and Lings, 2008: 2). However, if acceptance occurred there are positive emotions that result with a potential for increase respectively to the increase in the number of individuals that understand the process of that change and/or quality of their understanding. Among these positive emotions is a feeling of accomplishment that derives from the excitement of being involved in a change process and getting to achieve new things (Fullan, 2001: 8). The innovative CAIAT software of the present project has this sort of potential which is very likely to work as a catalyst to adopting it and gaining its benefit on a personal level.

3.2.2.1 Resistance to Change

According to Dalin (1978, As cited in Brown, 2003: 31-33), values, power, emotions, and practices (skills concerns) all become sources of resistance to change if corresponding characteristics of the individual appeared to be affected by that change. Douglas Barnes studied the perceptions of teachers and those of academics and educational advisory about an attempt for change, and found that while developers perceived resistance to change, teachers were worried about lack of resources which cause them, when they are introduced to new demanding developmental tasks, to complain about a shortage of resources (Russell and Munby, 1992: 14).

Teachers' conceptions affect change attempts. A good illustration of this impact is the movement of 'Language Across Curriculum,' which began in the London Association for the teaching of English, as a result of the report 'A Language for Life,' written by the Bullock Committee. It was set up by the British government in 1975, to look into the teaching of the English language and to invite schools to teach English reading, writing and talking through the different subject matter, though not necessarily English lessons, since English is used by students in those subjects. Consequently, all teachers are required to take the responsibility of English ‘curriculum’ within their teaching, in both primary and secondary schools, which mostly would require them to

make “radical changes” in their teaching practices. Although there was official support represented by different bodies (such as the London Association, LATE, and OFSTED,) unfortunately, the movement did not continue in secondary schools ten years later. Among the different factors that contributed to this end was the way teachers looked to the concept of subjects department from different dimensions (Russell and Munby, 1992: 27-30):

> In one very obvious way, the failure of 'Language Across the Curriculum' relates to the organisation of the curriculum not only in the [schools,] but in universities and elsewhere…Nevertheless, a price has to be paid for such a divided institution, and one loss is the weakness of any project that is intended to cross departmental boundaries…From the point of view of many teachers in secondary schools, language was by definition the responsibility of English specialists and therefore not of other departments. (Russell and Munby, 1992: 28-29)

Morrison (1998) elaborated on the use of pressure to induce change, criticising its implementation. He cited Kanji and Asher's (1993) seven suggestions of this pressure: top-down, bottom-up, peers, customers, competitors, performance data, and the set of required targets. He warns of what Senge (1990) noted: “the harder one pushes, the harder the system pushes back,” and highlighted that “it is more fruitful to remove the factors that are impeding it,” which is in line with what Everard and Morris (1990) stressed: that a “win/win” situation is a virtuous circle which apparently favoured the “win/lose” situation which ends up mostly in a “lose/lose” situation. Furthermore, Kanter (1983) warned that using force would reshape change as a threat to those concerned. Rather it should be introduced as an opportunity, for which motivation could play a significant role (Morrison, 1998: 129).

3.2.2.2 Motivation to Change

Joan Dean explained a number of teachers' motivators to change; among these are recognition, praise, and encouragement; seeing pupils developing and learning; challenge to professional skill; and a chance to contribute and shine (Dean, 1991: 17). I think that the last three are built-in motivators within the change process. However, this depends on the nature of that change attempt. Obviously, the present project includes such potentials. David Trethowan listed number of motivating items that relate to school policy, status and positive motivators. The latter consists of the job itself, cash, delegated responsibility, advancement, achievement, recognition by managers, and

personal growth (ibid, 18). Again, the last four are considered by the present project's setting. Advancement refers to the progress that the teachers will accomplish where they will experience how the new skills move them from a level of performance to a following level. Achievement is represented by the ultimate level that the teachers supposed to reach as a result of training and on-going practice. Moreover, they can define their level of achievement in this respect by repeating the use of an item after being analysed and rewritten accordingly. A following evaluation of the item quality will reveal the level of achievement that the teacher has reached. This shows the extent to which assessment of achievement is included within the design of the new intervention. Furthermore, the following two theories represent motivation trend.

Expectancy theory suggests that people's involvement in change is a result of their positive expectations of personal benefit from that change attempt. Lawler (1991) identified some of these benefits, namely: financial, job security, promotion, training and development, public recognition, feeling of accomplishment, and more interesting work. Morrison raised the issue of planning for career development as a powerful incentive and motivational strategy (Morrison, 1998: 132).

Exchange theory is seen to be more micro-political, in which individuals will be motivated if there is something which they will get in return: it is a ‘give and take’ rule. It stresses a mutual obligation and trust in both parties, and is thought to explain the organisational micro-politics. Hoyle (1986) mentioned four elements of micro-politics: interests (status, promotion, etc.), interest sets (formal and informal cliques, cabals, and coalitions), strategies (bargaining, dividing and ruling, co-operation, etc.), and power. A good example of micro-politics is what Bowe et al.'s work suggests: that the introduction of the National Curriculum and local management of schools in the UK has resulted in a competition between departments for scarce resources, which significantly increased micro-politics (Morrison, 1998: 133).

And as Eraut (2004a) puts it, “… and even when there is very little real threat, people become very anxious.” He outlined that “cultural norms” have constrains on people's thinking about change where they consider many practices as “unproblematic.” There could be also subcultures represented by professions or factions that could have their micro-political role. The political administrative factors could work for or against change, while individuals or groups could use their influence or opportunities available to them for opposing change that affect their power or increase their workload (Eraut, 2004a: 113). Joan Dean (1991: 8) supports that school culture represented by

assessment of attitudes is a key factor for effective PD. In terms of power, Hoyle (1986) explained how exchange theory underpins micro-politics through an example of head-teacher-teacher relationship:

> There is no locus of power. For example, whilst head-teachers have many properties to exchange, influencing, for example, material resources, promotion, esteem, autonomy, application of rules, teachers, too, have valuables that can be exchanges, for example, esteem and support for the head-teacher, opinion leadership, conformity, reputation (Morrison, 1998: 133).

Similarly, Morrison highlighted power characteristics in this regard that 'we might regard the process and dynamics of change [...] as a series of trade-offs (Morrison, 1998: 133). Power trade-offs and the other micro-politics should be considered in planning for change process so as not to cause a failure to the innovation's implementation. I believe that the two theories (Exchange and Expectancy) could integrate each other since each has got its underpinning concept from the diversity of people's interests. My design of the project appreciates the notion of teachers as intelligent, rational, and willing to adopt change, which is stressed by the rational-empirical strategy suggested by Dalin (1978), and at the same time targets the authority decision as an end that should then employ power to disseminate the new practice and provide a sustained application over a good period of time until it becomes institutionalised throughout the system.

This study's design considers the personal dimension for change, which could be seen in many aspects. First, it appreciates the issue of training and the PD so as to make the needed knowledge available to teachers. Second, it pays a high degree of attention to removing any potential sources of threats. Third, it appreciates Fullan's view of positive feelings as motivators to adopting change. Especially the feeling of accomplishment that derives from the excitement of being involved in a changing process and getting to achieve new things is considered important. The excitement of using a newly designed software and the excitement which accompany introducing a teacher to the concept and techniques of IAT, in which he/she becomes able to discover new information about him/herself is highly likely to encourage even more teachers to hunger to learn more and try furthering their options. This is congruent with expectancy theory.

In terms of exchange theory, the project design is initiated upon mutual trust between the two parties since it relies mainly on the teacher's belief in the extent of personal benefit to her/his PD. It provides them with benefits/opportunities, such as the

power of knowledge that comes from further understanding each individual's self-level of skills and developing them; being experts in a new innovation which opens up possibilities to be assigned to new roles in the future; and gaining training which accumulates to enrich a teacher's record.

3.3 Professional Development

According to Clift and Imrie (1981), most teachers prefer to conduct course appraisals by themselves, even though they can do this with the support of outsiders (Clift and Imrie, 1981, pp. 117-118). Therefore, the professional development (PD) of teachers is an important paradigm of this project's reflection on teaching because it implies that teachers take responsibility for self-learning and undertake different individualistic activities, such as action research. This all is congruent with their preferences, according to what Clift and Imrie reveal. Elements of PD include: building common knowledge and concepts; shared visions; a change in values and beliefs; translating new values into behaviours; and the systematic management of the resulting changes. Moreover, "Effective staff development will move professional staff from *what is*, to *what should be*. The key term is *change*, not change for the sake of change, but rather change for improved education" (Fitch & Kopp, 1990, pp. 4-5). It is evident that a significant investment in professional time, money, and new ideas is a result of PD (ibid, 1990, p. 36).

Dean (1991, pp. 4-5) listed a number of definitions covering PD, revealing that "the term 'professional development' suggests a process whereby teachers become more professional." She concluded that "a profession is an occupation which requires long training, involves theory as a background to practice, has its own code of behaviour and has a high degree of autonomy." She also indicated that PD includes enabling the individual to gain new knowledge, skills, and understanding which are gradually internalised to become part of her/his professional personality that is available any time (Dean, 1991, p. 19). Guskey defined a teacher's PD as "those processes and activities designed to enhance the professional knowledge, skills and attitudes of educators so that they might, in turn, improve the learning of students." It is also characterised as being intentional, on-going, and systematic (Guskey, 2000, p. 16). This particular definition mentions the impact on pupils' learning as a possibility rather than a certain consequence; however, the two definitions, as well as the majority of other definitions,

include three common elements that Ferreira and Santos et al. (2007) consider as being required for any kind of occupation, namely knowledge, skills/abilities, and attitudes.

In her paper entitled "Against Professional Development," McWilliam (2002) took a different perspective about PD, believing that PD as it stands does not fulfil its definition. She highlighted that knowledge of PD scripts constructed by developers, those who provide PD, "constitute" the developees. Therefore, PD is "knowledge production and a system of power relations." She also cited Hobart's (1993) term "charming absurdities," which describes the situation in which the trainer imposes upon his audience that pre-designed guidelines fit a particular condition, and then the people respond. She used that term to describe other "absurdities" in PD activities such as how it is employed to produce an individual that keeps up with the demands of enterprise culture and market requirements. The status of psychology for explaining the behaviour of people is another criticised aspect of PD. She uses the term, cited by Nikolas Rose (1990), "psychologization of the mundane," which describes how everything in people's lives have become subject to psychological consultations. Also, the idea that everyone ought to be a leader is another 'charming absurdity' because "If everyone is a leader, who is left to be led?" Many other points were argued such as attacking the attitudes of insisting on using PowerPoint presentations, online learning, multimedia and other technological means of learning, dispelling the myths that "technology will deliver" and "students prefer virtual pedagogies over face to face."

Finally, she revealed her concerns that "the knowledge which counts as professional development, and the processes through which that development is supposed to occur, ought to be scrutinised more closely." I believe that the point is not PD as a concept that needs to be criticised, but the sort of practice impeded within PD. Some inappropriate common practices could reshape PD's identity to reveal negative insight about it, such as that of McWilliam. Although McWilliam's stand is debatable, I think that what she stresses in her argument relates to PD that is designed/imposed by those who are more knowledgeable. However, in the present project, PD is more of a *self-learning* process which is triggered by typical training applied for outlining the main concepts and skills. The major factor then is the scaffolding practice driven by the utilisation of the CAIAT and ending with a new learning. This eliminates many of McWilliam's concerns.

Joan Dean listed a number of activities, other than training, that can elicit professional learning; among these are action research, reflecting on one's own

performance, and taking on responsibility by being involved in decision-making (Dean, 1991, p. 20). Actually, these aspects are present in this research. In education, PD is a key aspect for promoting change and can play a significant role in this aspect. Wenglinsky (2000) illustrated that when teachers received PD, their pupils improved in their performance and were better than their colleagues by more than a full grade level, while pupils taught by teachers that received PD on HCD exceeded their colleagues by 40% of a grade level (as cited in Vrasidas and Glass, 2004, p. 2). On the other hand, the notion of teachers being blamed for poor learning standards as a consequence of teachers' low level of preparation and/or low quality of teaching (Robinson, 1995; as cited in Smits, Wang, Towers, Crichton, Field, and Tarr, 2005) indicates the importance and necessity of PD.

Eraut stresses frequently in most of his writings that "time for reflection or explicit learning is difficult to find" (Eraut, 2005a, p. 1). Moreover, Black & William (1998b) and Darling-Hammond, Ancess, & Falk (1995) revealed similar considerations (Penuel and Yarnall, 2005, p. 5); hence the demands of a busy, crowded workplace contribute mainly to decreasing the opportunity of learning for CPD (Eraut, 2004a, p. 114). Pelgrum (2001) indicated the importance of PD for adopting ICT in education. In his study about obstacles to the integration of ICT in lower secondary schools, covering 26 worldwide countries, he listed the average findings for these obstacles, as shown in Table 3.1. He highlighted that the top 10 (above 50%) consisted of a mixture of material and non-material aspects. It is evident from the table that the most non-material obstacles are the lack of sufficient ICT knowledge and skills on the part of teachers, difficulty integrating ICT into instruction, insufficient time for students and for teachers, and the lack of supervisory and technical staff. This shed some light on the *soft requirements* that change-implementers need to consider when introducing ICT solutions into a school.

Table 3.1: Sorted list of obstacles to realising ICT-related goals, as perceived by educational practitioners across 26 countries

List of obstacles sorted by average percentage respondents across countries

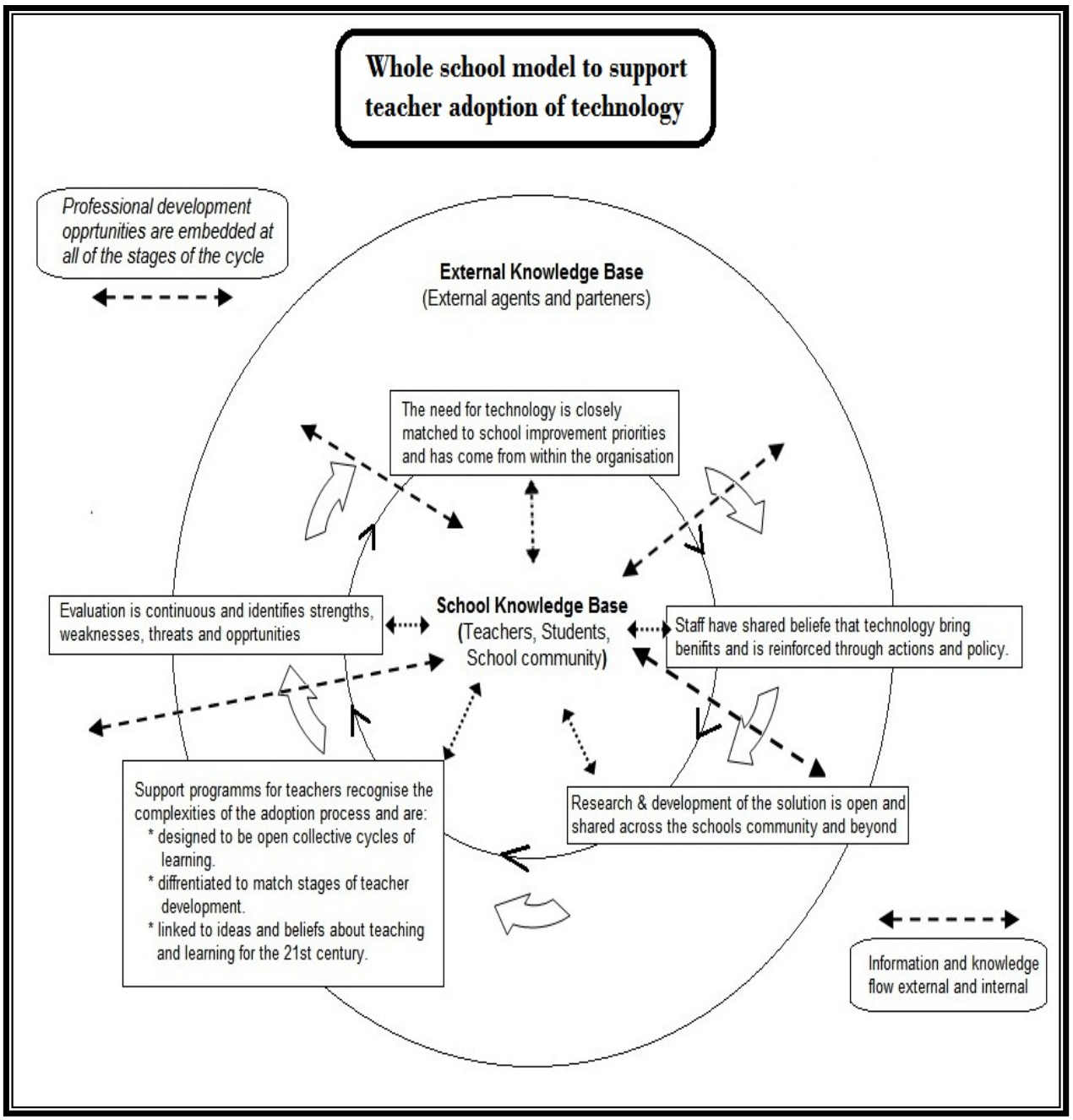
Obstacle	%	Obstacle	%
Insufficient number of computers	70	Quality teacher training too low	31
Teachers lack knowledge/skills	66	Software not adaptable enough	29
Difficult to integrate in instruction	58	Stud. know more than teachers	29
Scheduling comp. time	58	WWW: slow network performance	28
Insufficient peripherals	57	Lack of interest of teachers	27
Not enough copies of software	54	Difficult use low achieving studies	22
Insufficient teacher time	54	Telecom infrastructure weak	21
WWW: not enough simultaneous access	53	WWW: difficult finding information	21
Not enough supervision staff	52	WWW: information overload	20
Lack of technical assistance	51	Software curriculum incompatible	19
Outdated local school network	49	Lack administrative assistance	19
Not enough training opportunities	43	Software not in language instruction	18
WWW: no time teaching axplores	41	Lack support school board	17
WWW: no time school schedule	41	No plan prevent theft/vandalism	15
Lack info about software	38	Software culturally incompatible	12
WWW: not enough connections	35	Software too complicated to use	10
WWW: Insuffic. techn. support	34	Materials WWW poor quality	9
Not enough space to locate	32	WWW: complicated to connect	8
Weak infrastructure (telecommunications, etc.)	32	WWW: mail baskets overload	4

Eraut considers that an important requirement for achieving significant progress is continuity and external support which should take place for a considerable period of time (Eraut, 2005a: 1). Joan Dean (1991: 7, 10) indicated the role of the school where teachers act more positively if PD is part of the school and highlighted that it can be seen from the needs of individuals. The following example is to show the extent to which continuity and external support are important for the viability of a change attempt.

Betsworth (2003) has designed a model for whole school adoption of technology and implemented it in the form of case studies for two schools in two countries selected from the OECD project, Quo Vademus? The Transformation of Schooling in a Networked World (OECD 2002). The schools were chosen as representing diverse pairing; Crocodile Valley secondary school in an urban district in Canada, which has 1200 pupils made up of a diverse mix of need, ability, and ethnicity; and a municipal primary and lower secondary school in a provincial town in Denmark, which has 450 pupils with an even mix of gender. The model as shown in Figure 3.5 is a cyclic open sphere model that has school knowledge base at the centre and four assumptions flowing around and leaving freedom to two external factors to travel inward and

outward, namely: PD opportunities and information and knowledge. The four assumptions represent the heart of this model employment for examining an existing attempt at transformation. These are: the need of technology as coming from within the organisation, shared belief of staff about the benefits of technology, open and shared research and development (R&D) practices, open and relevant support programs for teachers and continuous evaluation. The researcher used these assumptions as instrumental guidelines for investigating each case study to find out the extent to which teachers' adoption of the new technology was successful and viable. The findings show that the second case in Denmark met all assumption of the model, and was thus considered successful; while the first case in Canada, although succeeded in implementation at the first years, failed to pursue this success after the head-teacher left when only 50% of the remaining teachers used technology to support their teaching compared to 75% at the time of before his departure. The investigation found that all four assumptions were met in the implementation except evaluation, because it has not indicated the possible threats to development overtime. Furthermore, the management and the coordinator that was employed for this purpose did not consider the threat of their possible departure from the school, which also combined with "the fact that the reform was to a large extent orchestrated by an external party." This party is the district which mandated this attempt without considering staff participation in decision-making. Moreover, the district has moved the innovative teachers that were acting as change agents from this school to other schools that will be applying a similar exercise (Betsworth, 2003: 37-56). I think that the major component of the benefited lesson from this experience is that change should not be applied from outside but is better to be seeded from inside, participation of individuals is required for this respect and evaluation and persistence of management support are essential for sustainable change.

Figure 3.5: Whole school model to support teacher adoption of technology (Betsworth, 2003: 56)



Teachers' decisions represent another dimension of their professional development course where their confidence and skills develop in this respect through reflective practice as they acknowledge that learning is the key to CPD, and furthermore, "maybe they don't solve all of the problems they confront, and maybe they make mistakes, but they never stop trying" (Henderson, 1992: 2). McNergney and Carrier (1980) pointed out the causal relationship between teachers' performance and students' learning explaining that "as a teacher's comfort level increases over time, he or

she becomes more concerned about understanding how the learner is affected by the innovation" (ibid, 1980: 151). Beyth-Marom and Dekel's (1985) definition of uncertainty-feeling states that "we feel uncertain about a certain question if we cannot provide for it one answer towards which we feel full confidence or complete faith." Floden and Clark believe that teachers' professional uncertainty decreases with experience and considers teaching as "evidently and inevitably uncertain," which Gujski and Ben-Peretz described as a "built-in uncertainty." To Eisner (1992), education is to learn how to deal with uncertainty (Gujski and Ben-Peretz, 2005).

Since ICT is utilised mainly in this project as a means for PD, it is worthwhile to indicate to the opponents of using ICT in education as a balanced view of the essence of this research concept. A number of writers have criticised the use of ICT in education (Postman, 1992; Oppenheimer, 1997; Becker, 2000; Bowers, 2000; Cuban, 2001; Warnick, 2001; Clare, 2005; Ferguson, 2005.) However, their criticism is targeted primarily at the use of ICT in learning situations and hence the direct impact on pupils. Bowers (2000), for example, who has authored a number of books in this area (David: 2007), stresses in most of his critiques justifications that computers make moral, political, cultural, and environment arguments. The cultural issue seems to be pivotal in his discussions.

Most if not all writers though, indicated findings that reveal no significance of using ICT for low performing pupils. Furthermore, they reported a negative impact. Clare (2005) wrote about how a Royal Economic Society study highlighted that "computer use in schools does not seem to contribute substantially to students' learning of basic skills such as maths or reading." He explained that the low educational achievement of pupils of schools "generously equipped with computers" is because the effective teacher's instruction disappeared as a result of the computer-based instruction (Clare: 2005.) Cuban (2001) explained that in the studied schools, he found "no substantial evidence of students increasing their academic achievement as a result of using information technologies."

Interestingly, he commented on what he called the "blame and train" approach, which describes a common practice with technophobic teachers forced to train, whereby he found that many of the teachers were actually learning the computer and using it extensively. However, their use was not primarily for instructing students but to "prepare their work, communicate with parents, colleagues and students, maintain records, and carry out research." Those who utilised computer for instruction stood at

less than 5%. To the present research, this is very relevant because it shows that teachers' interest in utilising ICT for their work, such as self-learning and reflection, which it is suggested the CAIAT will stimulate, is at a substantially higher level of compared to their interest in using it for instruction. The criticism of using ICT in pedagogical practices does not apply to the purpose of the present work, in which the teacher is the prime goal, not the pupil.

I should highlight that the present project has been designed to meet most of the preceding PD considerations in terms of time allowed for PD, continuity, the internalisation of skills and knowledge, reflection, problem-solving, decision-making, school and/or district support, and targeting pupils' learning as a potential goal. However, I will elaborate next on two major PD themes that will be employed practically within the present process.

3.3.1 Professional Learning and Reflection

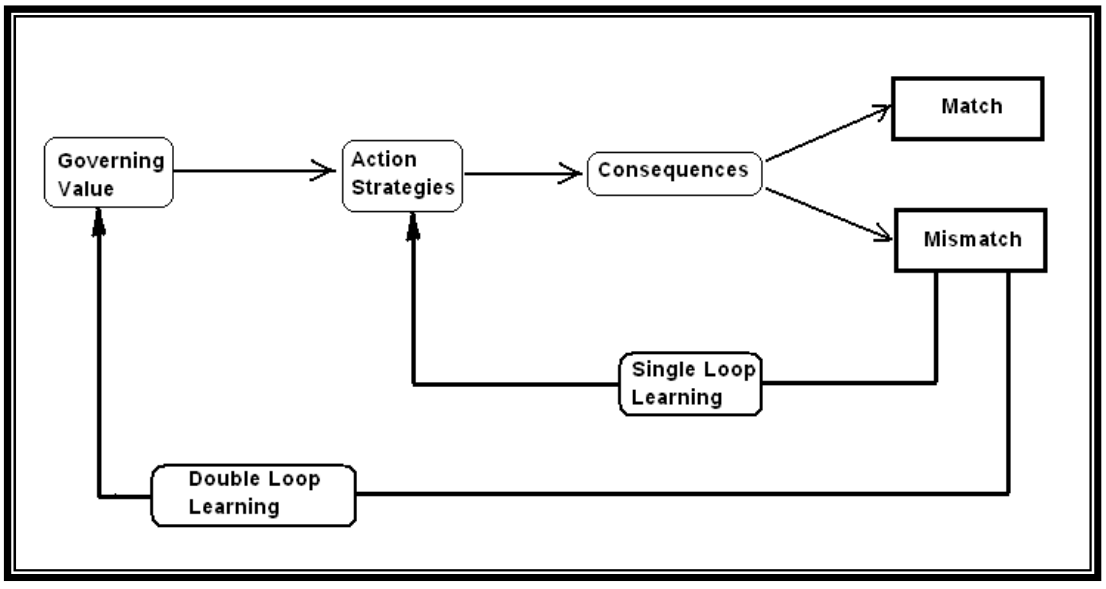
Reflection may appear in the literature of PD in different phrases such as: reflective judgement, reflective practice, reflective thinking, critical reflection, reasoning, inquiry, problem-solving, thinking or reviewing (Kitchener, 1983 in Moon, 1999). Reflection is the core aspect of PD since it represents the major outcome. The meaning of reflection as seen by Eraut (2004b: 47) is "the action of turning (back) or fixing our thoughts on some subject; mediation, deep or serious consideration." Since teachers were students for almost 20 years, their background as learners is very likely to shape their teaching styles and practices. This has been shown by relevant research (Ball, 1990 and Lortie, 1975 in Vrasidas and Glass, 2004: 3). Vrasidas and Glass (2004) denote the fact that what is stressed by literature about pupils' learning and the different meaningful activities that is suggested for eliciting their optimum learning capabilities applies for their teachers as well in terms of PD discourse. They believe that optimum PD is the one that results in making them think and act like 'experts,' and consider that the circularity of the teaching profession that comes from being 'taught' how to 'teach' distinguishes it from other occupations (Vrasidas and Glass, 2004: 2-3). Eraut highlights his concept of inter-related learning trajectories as follows:

> inter-related learning trajectories…does not assume that learning follows stages that correspond to artificial stopping points associated with competence or qualifications [and does not] necessarily progress onwards and upwards (Eraut, 2005a: 1).

He thinks that skills development in learning to learn and critical thinking need continuity and coordination in which a small number of key courses use suitable content for skills development. In this respect, he advised to start small but long, where progression in the areas of PD should be mapped onto relevant learning trajectories which could develop the "learner's self-directed learning and sense of urgency" (Eraut, 2005a: 1).

Argyris & Schon (1974) introduced the idea of single and double-loop learning, showing that better results in terms of reflection could be gained through double-loop learning in which the theory-in-use (which individuals adopt to explain their behaviour) is to be abandoned, the individuals' assumptions to be challenged, and a wide range of evidence to be discussed. This all should result in reframing the problems so new solutions can be developed (Eraut, 2004b: 51). Figure 3.6 shows a model of organisational learning in which a single-loop learning path is undertaken when action strategies only need correction, whereas a double-loop learning path is followed when alteration of action strategies is not enough and, further, there is a need to alter values or variables that govern these strategies (Argyris & Schon, 1978: 262-263).

Figure 3.6: Organisational learning model (Argyris & Schon, 1978)



I have to highlight that "Theory-in-Use" is one of the action theories which is used to produce the action. It is found to be dominant and free of different characteristics such as gender, race, education, social status, wealth, type, age, size of organisation, and culture. The consequences of actions upon this theory are devastating

in which “escalating errors” is not the only continuous aspect and hence this would produce a “defensive reasoning” mind-set which in turn produces a set of skills that are against self-learning and of “systematic denial.” Furthermore, this could produce organisational defensive routines that are against learning since they aim to reason for embarrassment avoidance and are “overprotective.” (Argyris & Schon, 1978: 264-266).

Eraut's research has shown that explicit and implicit learning are more likely within group activities, tackling challenging tasks and problem solving (Eraut et al. 2004a,b in Eraut, 2005: 1). There are four suggested learning styles for PD: assimilator, detail learner, passive learner, and shaper. The latter is thought to be the one to get the most benefit since he/she creates new structures for the new learning (Dean, 1991: 20).

Studies tend to show that the impact of in-service training is minimal since teachers mostly do not apply what they have learnt and thus put the responsibility on two parties: in-service trainers and school management (Dean , 1991: 20). Eraut (2005b: 61) considers that “a profession is better understood as an applied field rather than a discipline, because its rational derives from its social purpose and not from any distinctive form of knowledge.” He argues that interaction between theory and practice engenders the “personal theory” that practitioners construct as a result of their empirical research, their maxims, preferred ideologies of the profession, and views of society. He classifies evidence-based practice as a theory of practice which creates forms of explanations of observed practices and contributes then to the understanding of these practices. Eraut points primarily to reflection and review process that links research evidence with practice in order to reap a better PD learning. Teachers need to have their time for reflection and reviewing what they have performed in all areas of their profession, be it in their pedagogy, classroom administration, or assessment. Methodologically, this is embodied in action research.

3.3.2 Action Research: A methodological agent for PD

Bybee and Loucks-Horsley (2000) believe that action research (AR) is the strategy that helps teachers to learn and gives them tools for “less formalised” learning, and they explained that policies and standards, though necessary for enabling change, are not enough. In this sense, PD provides a “critical companion” for the sake of good implementation. AR is seen as a means of extending teachers' professional autonomy (Winch and Forman-Peck, 2000: 165). Cohen, Manion, and Morrison (2000) define AR as “a small-scale intervention in the functioning in the real world and a close

examination of the effects of such an intervention." By means of AR, the gap between research and practice could be filled, social justice could be maximised, and transformations of individual, culture of groups, institutions or societies could be undertaken. The scope of AR application is very wide and covers most needs of teacher development. It also covers a variety of areas; among those are evaluative procedures, modifying attitudes or values, improving teaching methods, and continuing professional development of teachers which is the interest of the present research's implementation of AR. It is worthwhile to point out Kemmis and McTaggart's (1992) identification of AR: "It is not research done on other people, action research is research by particular people on their own work, to help them improve what they do" (Cohen et al., 2000: 226-227). Therefore, I see AR as a tool for teachers to learn and/or appreciate any new educational technique. They may be trained well on applying something, but AR will help them to see the merits and distinct aspects of that new practice.

Carr & Kemis (1986) classified AR into three models: technical, practical/deliberative, and critical. These models interpret each of the four steps of AR suggested by Lewin - reconnaissance, plan, implementation, and evaluation in terms of the three central concepts of AR: action, research, and evaluation. Each model has its own aim, where effective practice is the aim of the technical, increased understanding is the aim of the practical/deliberative, and empowering individuals to change their social life is the aim of the critical (Winch and Forman-Peck, 2000: 167).

North (1987) considers AR a kind of "practice-as-inquiry," in which teachers, after identifying a problem, would search for solutions, test them, and validate their observations to disseminate what they found. This vision of AR may require it to follow a traditional quantitative method, while Donald Schon's (1983, 1987) aim of AR is to inform and change on-going practices as the teacher "reflects both while engaged in action and subsequently on the action itself" and this trend sounds biasing to the qualitative approach. However, Newman (2000) pointed out the flexibility of AR as a new "interesting form of practice-as-inquiry" for which there is no one 'right' way of doing it (Newman, 2000).

This research will focus on training teachers and imply a course of continuous learning, within which teachers meet and work co-operatively, and therefore be expected to follow an AR activity to fulfil the project objectives. They will also work individually to self-learn the project's different skills, which entails that they again may

follow an action learning individually. The link between learning and action, as the approach suggests, helps to gain effective and efficient implementation (Rondinelli 1983a; 1983b in Rondinelli et al. 1990: 18).

3.4 Diffusion of Innovation

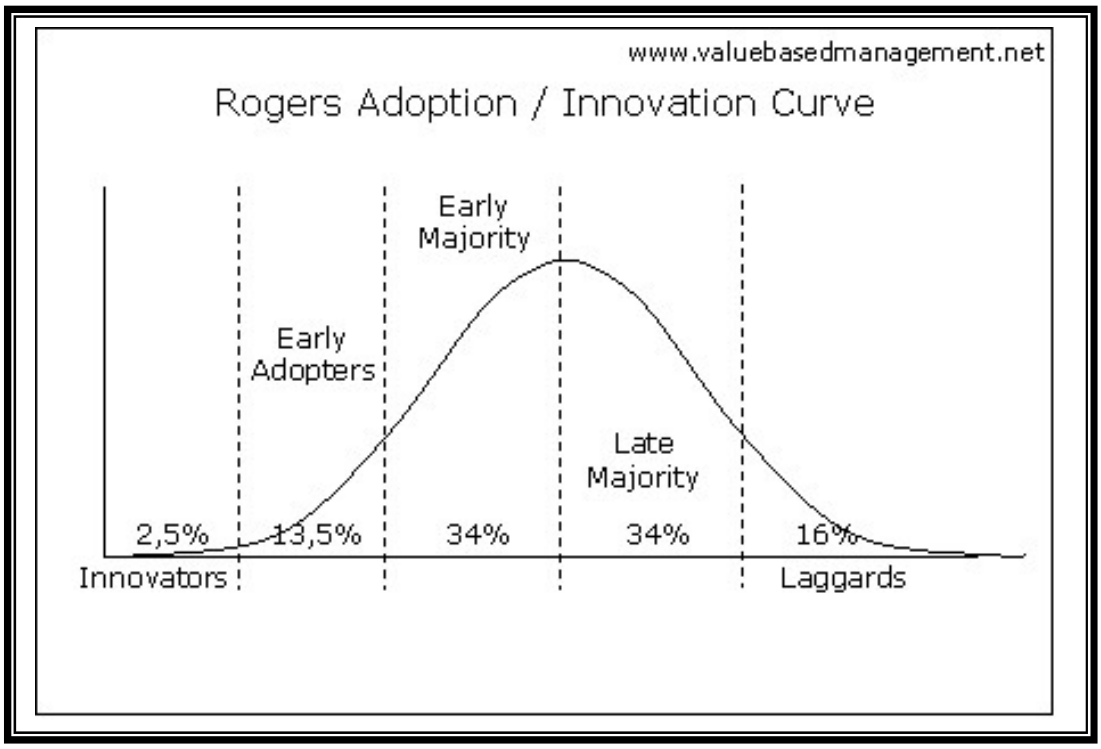
In 1962 Everett Rogers published his theory of Diffusion of Innovation (DoI) which describes how the adoption of new innovation occurs. It is widely used as a theoretical framework in information technology (IT) research, (Roger, 1999 and Rabina and Walczyk, 2007). However, it also offers a conceptual framework at a global level not yet concerned with IT exclusively, (Dillon and Morris, 1996: 5). This theory is thought to be a good start on which planners may draw for their new innovation applications. Dillon and Morris (1996) believe that the different suggestions of innovation adoption frameworks are part of this theory (Rabina and Walczyk, 2007). They have highlighted the role of “user acceptance” believing that "lack of user acceptance is a significant impediment to the success of new information systems (Dillon and Morris, 1996: 3).

I consider that the present project is a seed for diffusing HCD/CAIAT intervention to a wider scale, thus its design should comply with DoI considerations, especially the user acceptance notion. It is anticipated that many of the teachers could experience an embarrassment by their poor performance as revealed by the CAIAT outcomes. This is the very reason that invited me to design the project on a way that does not violate the individual's privacy therefore it is exercised as a self PD practice. In doing so, I am intending to accomplish user acceptance of the new innovation as a pivot for the success of accepting the CAIAT intervention and also to follow the *personal power* path of change (Figure 3.2). The innovation adoption curve that appears in Diagram 3.1 represents a pattern of DoI which classifies people's responses to innovation into five categories:

- Innovators: They are brave enough to pull change.
- Early Adopters: Try out the new ideas in a careful way.
- Early Majority: Thoughtful people but accepting change more quickly than the average.
- Late Majority: Sceptical people that only adopt new ideas when the majority does.

- Laggards: Traditional people, who are critical towards new ideas or maybe socially isolate. Both will only take the new attitude if it has become a mainstream.

Diagram 3.1: Innovation adoption curve by Rogers (Value Based Management.net, 2004)



From the curve, one can see that early majority versus the late majority represent the biggest range of the diffusion which is 68% for both. Innovators and the early adopters are the leaders in terms of time scale, which calls for the innovation appliers to focus on these two at the beginning in order to let them take the role of change agents at a later stage. To facilitate this functional task, Rogers has illustrated psychological and social characteristics of each of these categories' individuals. (See: Dillon and Morris, 1996, Roger 1999 and Rabina and Walczyk, 2007). Rogers curve is comparable to the S-curve of innovation's growth of revenue which is shown at Diagram 3.2. S-curve could take different shapes according to the slope of the curve which depends on the rate of diffusion: with a rapid rate, the slope will be very steep and vice versa. (Mahajan and Peterson, 1985: 9). Diagram 3.3 shows how the Rogers Bell-curve, which represents response/adoption, is reflected by S-curve of innovation, which reflects growth of revenue.

Diagram 3.2: S-curve for growth of innovation diffusion (Wikipedia[19], 2009)

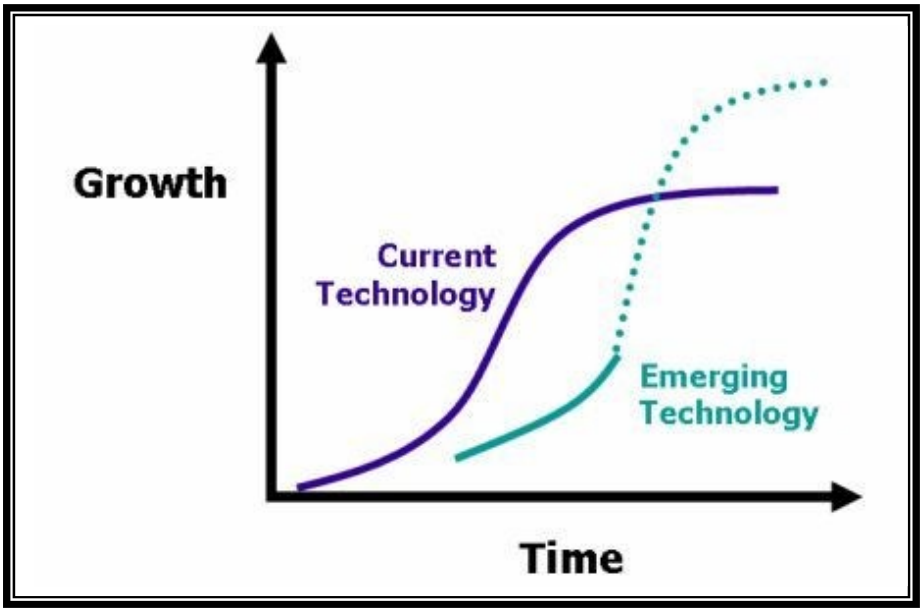
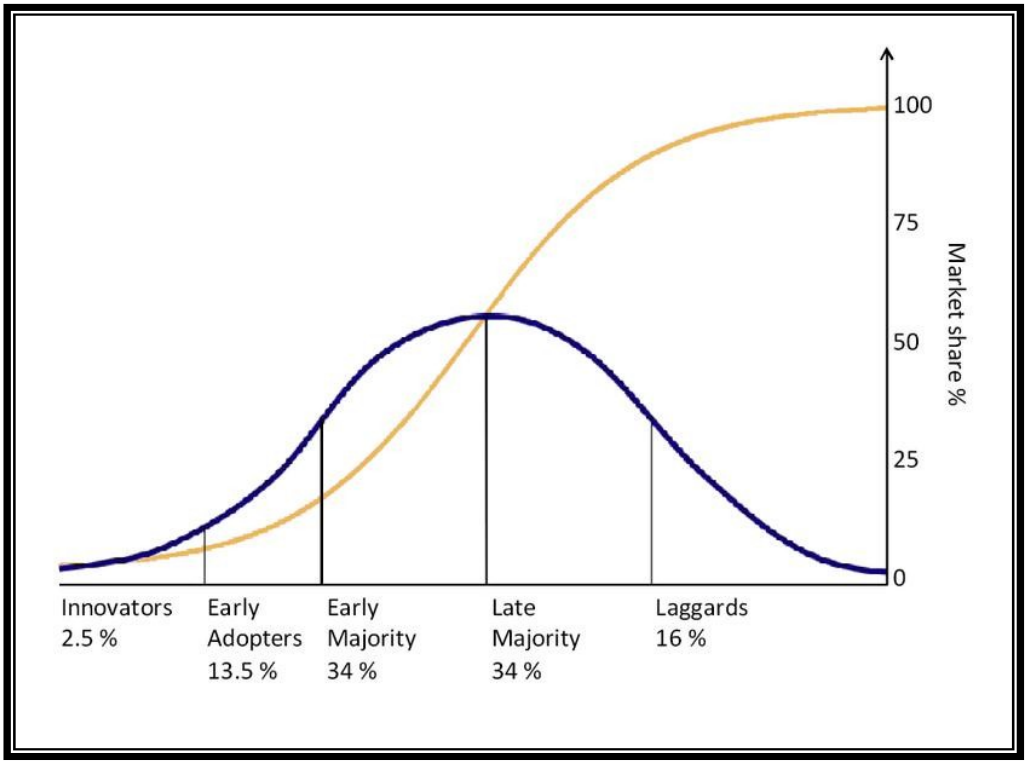


Diagram 3.3: S-curve compared to Bell-curve (Wikipedia, 2009)



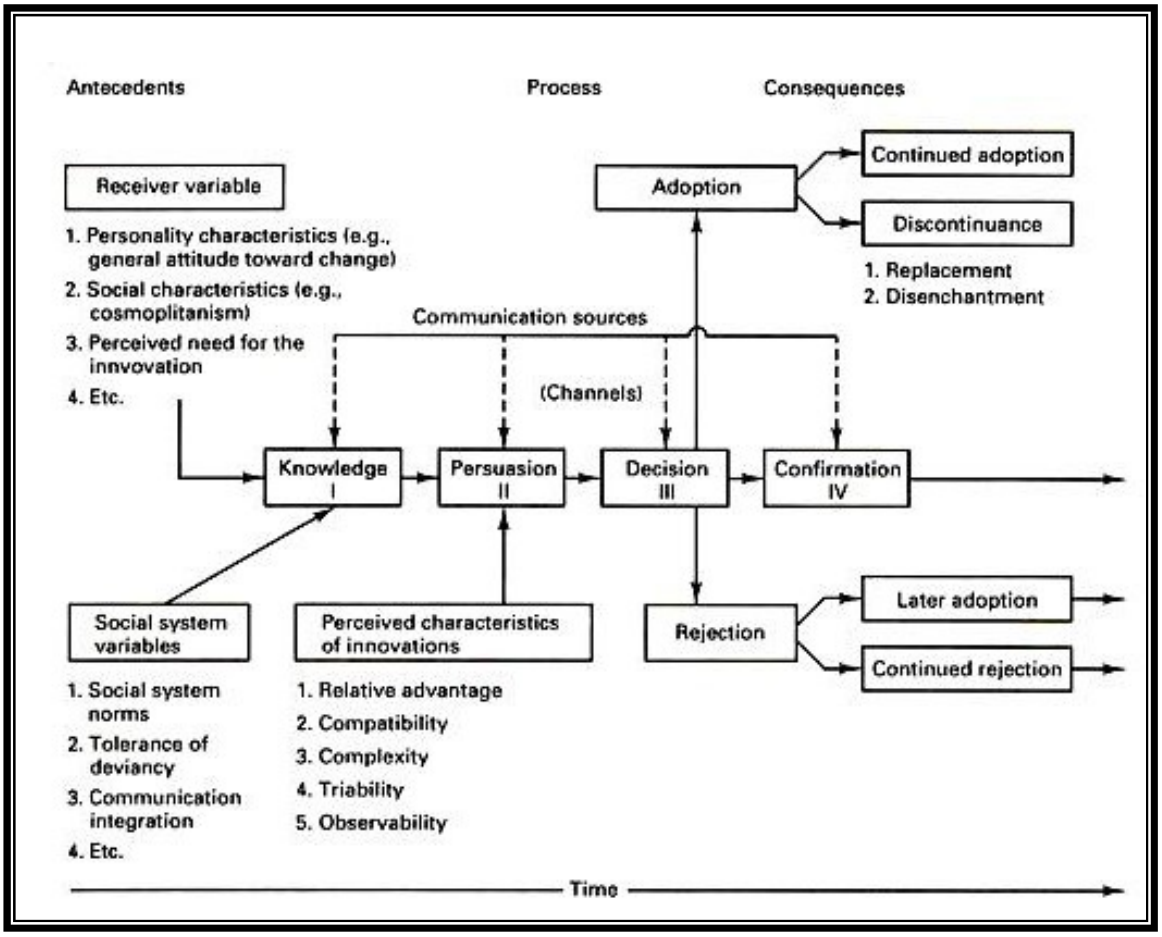
Rogers categorisation expresses DoI in terms of responses while DoI has five stages in terms of tasks: knowledge, persuasion, decision, implementation, and confirmation -- which include reinforcement as a result of positive outcome (Roger, 1999). The first two components, knowledge and persuasion, are reflected by the first two stages of the present project's model of *personal power* (Figure 3.2). Figure 3.7

illustrates the flow of the five stages as well as variables and characteristics that interact with or control the process. The item of interest is 'decision' where which 'adoption' takes two possibilities, rejection (the downward route) or adoption (the upward route). Continual adoption is the ultimate goal of the present project because it is aimed that using the CAIAT becomes part of the teachers' on-going practice. However, measuring this long-term effect is beyond the limits of the research; but, it is possible to measure early indicators to it. Among these are: the effectiveness of building the new innovation's skills and the early adoption of the innovation as measured in this project's design hence I have classified the project's outcomes into two dimensions: *effectiveness* and *adoption*. These will be explained further in Chapter 5. Moreover, I think that the route of rejection could provide us with yet another insight. The sub-rout of ''continued rejection" could be logically abandoned if we found out that the innovation was adopted from the beginning. Thereafter, later adoption could be looked at as a next degree in this respect.

Another important aspect that this theory conceptually provided to the present project is the perceived characteristics of innovation as a catalyst factor for persuasion. These are: Relative advantage, compatibility, complexity, trialability, and observability (Roger, 1999). These characteristics are taken into account within the CAIAT software wherein the *relative advantage* is obviously its core purpose. *Compatibility* is present by the fact that this software is, in general, similar to the most common software packages that teachers use (such as MS Word©, Excel©). Also, *compatibility* is seen in the congruency between the CAIAT functions, processes and outcomes with those of the manual IAT. A similar software package, ITEMAN©, as will be explained later, does not have this sort of compatibility. *Complexity* is seen in the way the CAIAT can work with both kinds of questions: closed and open-ended questions, as well as different methods of entering data to facilitate to the user dealing with the data and to make all possibilities available. *Trialability* is mostly available in any innovation that relates to ICT since the user will be able to try it out safely as many times as could be imagined and also could do different copies for trials while altering between these very easily. *Observability* is limited in this project's design because adoption could not be easily tracked or followed upon. However, observation has been tackled during the project by different data collection procedures: observation during workshops, during lessons, and case study. The preceding elaboration on how do these five factors exist in the CAIAT

is to identify the extent in which the present project's inclusions and design comply with DoI theory's implication.

Figure 3.7 Conceptual model of DoI theory
Source: Rogers (1995). In Theory Cluster (2004)



3.5 Change in KSA

In recent years, the Saudi educational context has responded positively to attempts at change that have taken place in various areas of education in general and in areas related to ICT specifically. Examples of these efforts at the ministry level are the successful computerised school systems, which manage pupils' marks, and teacher transfers and registration online. In the Al-Ahsa area, where this research project was undertaken and where I have a first-hand experience, there are various further examples, such as electronic attendance scanners; the utilisation of email for official administrative communication instead of hardcopy letters; online questionnaires; and registration portals for training, employment and computer illiteracy programs. This is in parallel to

a number of applications being used here and there according to the particular needs of each school, section or department. These may include dealing with stock control or financial affairs (as in the school cafeteria), or record-keeping with scanners.

Moreover, different educational interventions, not related to ICT, have been completed successfully, hence the required change has been reaped. Among these are *Comprehensive Evaluation*, which is similar in concept to OFSTED in the UK; *Focus Schools*, which involves focusing on specific schools in order to raise the standards; *Exchangeable–Evaluation*, which is a 360° model of employee evaluation, and other similar interventions that tackle change from various perspectives. These projects are between two and six years old, which indicates that they are approaching institutionalisation.

In terms of pedagogy, many such attempts have been implemented, with a good level of acceptance by all the beneficiaries; for example, there is the work of Abdulrahman Al-Garfi (2010), who carried out his project in a Saudi secondary school. His research set out to determine teachers' and pupils' perceptions of and responses to implementing cooperative learning methods. From the two male teachers involved in the study, along with their 39 pupils, Al-Garfi found that they reported many benefits and "aspects of their practice that illustrate a shift to a more pupil-centred classroom." Pupils revealed their "enjoyment" and the freedom and opportunity to be responsible for their mutual learning. The study recommended that cooperative learning be considered within both pre- and in-service training (Al-Garfi, 2010, p. 2).

Most notably, some national projects that were initiated by the MoE and applied in a number of LEAs have been evaluated and found feasible. Among these, I will present three major projects: Leading Schools, Teaching English in Primary School, and Merging Special Educational Needs (SEN) Pupils in State Schools. Leading schools are similar in concept to what is called in the U.S.*: Alternative Schools* or *Schools of Choice*. It focuses on non-classical styles of teaching such as cooperative learning and open learning environment and provides a new setting of curriculum provision so as it becomes more adaptable to students' interests and serves the goals of the school as set out in its initiation document (Al-Awwad et al., 2010a). It also borrows the concept of *Charter Schools* but with an accommodated pattern that suites the centralised Saudi educational system. In this respect, it includes the creation of a school board[20] led by someone other than the school head-teacher. The board's responsibility is to establish a charter document for the school which outlines all of its major aims and strategies to

provide qualitative education. It is also responsible for communicating the school's activities, initiatives, and inventions to the community. It could also participate in raising school funds and solving problems (Al-Khlaif, 2000). The curriculum of the school has some level of flexibility to meet the requirements of teaching by modern pedagogical trends such as cooperative learning, problem-solving and self-learning. There is a lesson period dedicated to extracurricular activities and the role of the pupils' counsellor is a very active one. A senior teacher[21] is also appointed for liaising amongst different teachers of the subject and also for planning and managing professional developmental programmes for teaching staff (Al-Owaisheq & Al-Suwailem, 2000). The project was applied in 24 schools distributed over seven major provinces of KSA including a total of 13,678 pupils.

Al-Harthy (2002), who led a team to design the curriculum requirements for Leading Schools, has concluded an evaluative survey on a limited sample to assess the opinions of the teaching and administrative staff of these schools about their experience within this change attempt. Responses revealed that 81% of the teachers and 75% of the administrative staff feel that they are happy with the new experience. In response to whether they have benefited, 75% of the teachers and 72% of the administrators revealed that their educational concepts have improved. Another evaluative national study by Al-Awwad et al. (2010a) has covered different leading schools with a larger sample and has shown that the trend has been successful. Following the CIPP model of program evaluation, the scope of that evaluative study has extended to include inputs, processes, and outputs of the project. Although some findings indicated some important shortcomings in the application, the overall findings revealed that the project was successful. By following the recommendations of the study, the implementation could be expanded to reach all Saudi schools instead of only the experimental 24 schools. Most relevant to the present study, one dimension of the findings was “acceptance of the new paradigm of *leading schools*” which showed that there were some obstacles that decreased the degree of acceptance. Examples of these are absence of incentives for teachers, given that there is an additional workload resulting from this school model; also, teachers felt that they were not part of the decision-making, which decreased their impetus in relation to some aspects of the project. The point is that none of these obstacles related to a negative attitude of the project's individuals but all related to a lack of support by the MoE. Summarising, this highlights the fact that this change

attempt found personal acceptance resulted in a number of its dimensions being applied successfully[22].

English was not part of the primary school curriculum before 2005, but then, as a result of a royal decree, dated Aug 2003, it was decided that English should be included within the sixth grade of primary school in a two-year period (Journal for Educational Documentation, 2003). Many opponents declared that society is against this movement, which called on the MoE to conduct an evaluative study to find out the extent to which this is a solid position across the whole community of the state. Al-Damigh & Al-Shumaimry (2010) conducted this study and found many important results. In addition to the study's findings about textbooks and levels of pupils' achievement, it assessed different stakeholders' opinions about the new approach. For pupils, they reported that almost 75% of the pupils agree to learn English from primary school. Moreover, more than third of the sample (almost 35%) wish that English be taught from the first grade of primary school. Pupils also reported that their parents are helping them at home in learning English, which reflects the parents' positive attitude. Actually, parents' responses to the survey have shown that 90% of them agree that English should be taught in primary school. More than 80% of English teachers and educational supervisors agree with this as well. All these figures reflect that there is a strongly positive attitude towards the new change albeit some voices in the community were raised against it. From my perspective, I interpret this from the fact that Saudi individuals have become more liberated in thinking about what is useful and what is not. They tend to decide in the light of what they find in practice not in the light of others' prejudices. This state of open-mindedness is a rich opportunity for future change attempts to take place in KSA both socially and educationally.

Merging SEN Pupils in state schools began in boys' schools in 1996 and has continued until today, encompassing boys and girls alike. The merging means that SEN children are mainstreamed with ordinary children in public schools instead of being educated in SEN institutes/schools only. This is applied with the provision of SEN needed services while they are in a state school. There are many benefits from such interventions, especially those relating to psychological/social attributes. Furthermore, the provision of education to SEN pupils will be more accessible, instead of having long waiting lists for SEN schools/institutes (Al-Mousa et al., 2008). However, many obstacles and difficulties could result in the failure of such a trend, because children of this type within state schools have many needs, especially those pertaining to

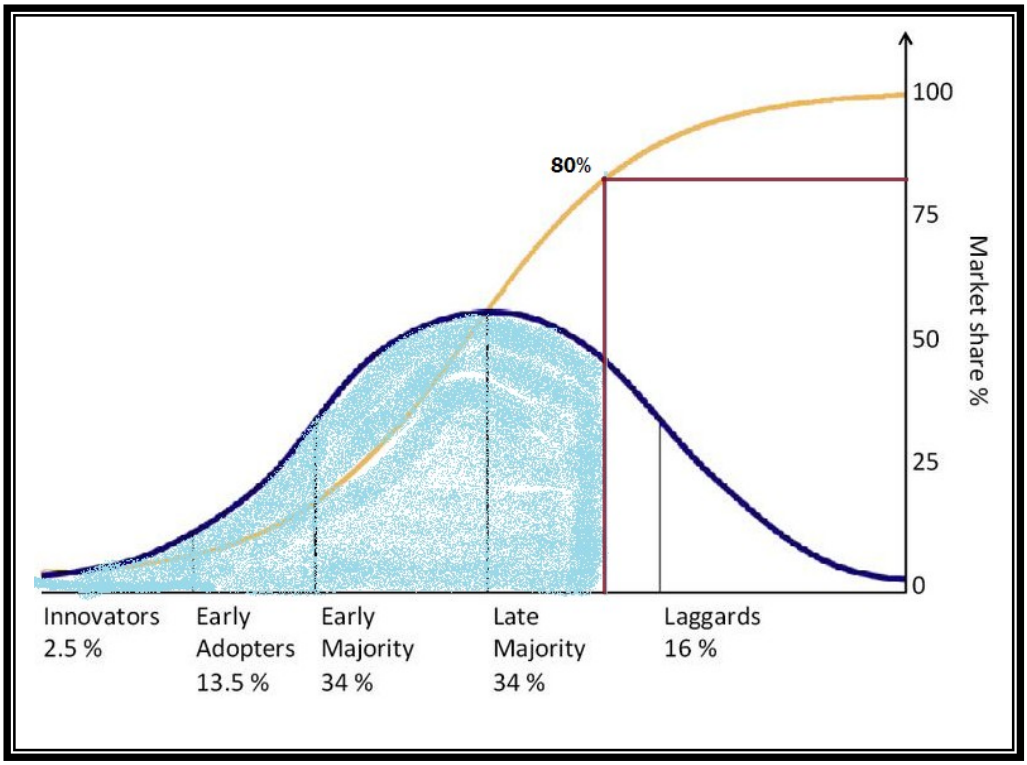
acceptance by all stakeholders such as state school teachers, pupils, staff and parents. The evaluative study which Al-Mousa et al. (2008, p. 536) conducted has investigated this concern by assessing the following major issues:

1. The effect of the educational environment (mainstreaming settings versus segregated settings) on the academic achievement, adaptive behaviour, and self-concept of students with special educational needs.
2. The attitudes of personnel in regular schools and in special education institutes, and the attitudes of normal students and their exceptional peers as well as their parents toward educational mainstreaming.
3. The positive and negative impact of mainstreaming on both home and school environments in the Kingdom.

For the first item, the study has found that mixing SEN pupils with their peers in state schools has no negative impact on their academic achievement; rather, there are improvements reported. For the second item, overall findings about attitudes of state school staff, pupils, parents, SEN pupils, and their parents are positive about this change; indeed, there is an indication that these positive attitudes are increasing over time. For the third item, the study reported both positive and negative impacts of the intervention, highlighting that the negative impacts relate to the lack of quality in application and not to the concept of the project or acceptance of its notion by stakeholders. This suggests that these negative impacts may be easily overcome in the future.

Most of the preceding studies have identified percentages between 70% and 90% for people's acceptance to the change attempt under investigation; therefore, an average point of 80% could be considered to conceptualise where these change attempts lie in the model of an S-curve. Thus, drawing this point on the curve of Diagram 3.3 results in the shape illustrated by Diagram 3.4 below. This shows that not only the first three groups of DoI theory adopt the new innovation but also a quite portion of the “late majority” joins from the beginning. The rest of these along with “late laggards” are very likely the main source of reported complaints from lack of characteristics of the innovation or requirements for ideal application.

Diagram 3.4: Average percentage of Saudi users' acceptance for some studied change attempts represented on the S-curve as compared to the Bell-curve



Chapter 4

Quality of Assessment

4.1 Assessment: International Concern for Quality

Assessment has a pivotal role in many contexts. One important role is investigating education quality. In UK for example, I cite the popular article of Black and William (1998) "Inside the Black Box," in which they concluded that improved formative assessment raises student achievement. They outline practical steps that fulfil what the article "raising standards through classroom assessment" calls for. In a following article (2004) entitled "Working Inside the Black Box: Assessment for Learning in the Classroom," they commented on their former paper's two questions. The first question was: "is there evidence that improving formative assessment raises standards?" and their answer was "an unequivocal yes, a conclusion based on a review of evidence published in over 250 articles by researchers from several countries." The second question was: "is there evidence that there is room for improvement?" and they highlighted a positive answer but faces three problems:

- The assessment methods that teachers use are not effective in promoting effective learning.
- Grading practices tend to emphasise competition rather than personal improvement.
- Assessment feedback often has a negative impact, particularly on low-achieving students (Black et al., 2004).

This well-known article and its echo throughout relevant literature illustrate the extent to which assessment is important for discussing the UK education. Furthermore, and on the official level, Google provided 557 links by searching the OFSTED website for the term "pupils' assessment" and 517 links for the term "teacher assessment." This indicates the level of interest that OFSTED attributes to the assessment issue throughout its reports. In USA, professional teacher organisations included assessment in the standards of the profession:

> Quality assessment … hinges on the process of setting up conditions so that the classroom, the school, [and] the community become centres of inquiry where students, teachers[,] and other members of the community investigate their own learning, both individually and collaboratively (NCTE, 1994, p. 7, as cited in Smits et al., 2005).

In KSA, the *pupil assessment document* provides a framework for assessment practices both for teachers and authorities. The instructional and learning-oriented purposes appear explicitly in two of the five goals of the document, namely 2 and 3:

> 2. To identify pupil's level of achievement and to recognise the extent to which he was able to fulfil aims and objectives of education strategy of KSA.
> 3. To provide both pupils and officials of the required information for improving learning as well as efficiency of curriculum and pedagogy (HCEP, 2006: 3).

The General Directorate of Educational Research (GDER) has announced its priority for research as represented by ten subjects. Among these, three are related to assessment namely (MoE website, 2009):

1. Preparation of national tests for five major subjects of all key stages.
2. Evaluation of continuous assessment experience in primary schooling.
3. Evaluation of decentralisation of high school national test.

This shows that the MoE has much interest in research that relates to assessment and thus reflects a concern of the ministry to improve and develop the quality and input of assessment.

4.2 Assessment Debates

Assessment is one of the educational subjects that has got a huge number of writings and research including discussions and debates. This is due to the different areas of interest within assessment and different approaches through which assessment could be employed for improving or developing education. To highlight this aspect, three of the major and common debates of assessment will be presented: criterion vs. norm referencing tests, external vs. internal (teacher-based) assessment, and assessing HCD skills.

4.2.1 Criterion vs. Norm Referencing Tests

In norm-referenced tests, pupils' performance is measured in relation to a "normative" group (e.g. a state, school, or class). In criterion-referenced tests, pupils' performance is measured in relation to a pre-defined criteria that are described in terms of the curriculum objectives or standards that needed to be mastered (Johnson, 1986: 95). The latter is concerned with measuring learning quantity and quality (Satterly, 1994: 57) whereas the former is more concerned with the level of discrimination available among pupils[23] (Freeman and Lewis, 1998: 16). Therefore, if a test is to select a group of pupils for a scholarship then a norm-referenced test should be constructed,

i.e. highly discriminating test items need to be written; but if the test is to discover whether the pupils have understood thoroughly what they had learnt, then a criterion-referenced test should be constructed. Criterion-referenced tests are considered as more 'informative' and of 'greater utility' for teaching than norm-referenced (Freeman and Lewis, 1998: 20), thus they are much more appropriate for the formative role of assessment from this point of view. However, some writers consider these 'slippery' concepts as Butterfield (1995: 117) pointed out. In practice, most tests contain an element of both.

> Tests constructed to produce norm-referenced outcomes can be regarded as criterion-referenced … Similarly, criterion-referenced instruments can be used to rank candidates and compare their performance and may therefore be treated as norm-referenced instruments … However the point stands that it is the use made of results, rather than the nature of items that most clearly distinguishes between test types (Lewin, 1997: 13-14).

Discrimination is a baseline for discussing differences between the two concepts. In norm-referencing, discrimination is a major characteristic of the optimum test, because the main function of the test is to show differences between candidates. It is expected therefore that optimum norm-referenced test is the one in which individuals are statistically distributed normally, i.e. on a Gaussian curve type as Diagram 4.1 depicts. By contrast, criterion-referenced tests do not aim to differentiate between candidates, but rather to measure the level of achievement they gained in terms of the educational objectives around which the test items constructed. When test results are distributed on an extreme distribution, such as the one shown in Diagram 4.2, then it would mean that the pupils have achieved most of the assigned objectives that were both taught and tested. Although a CR item is possible to discriminate around the zone where its outcome is achieved, e.g. a high score peak similar to the one on Diagram 4.2, this discrimination is not necessarily a quality characteristic of CR item.

Diagram 4.1: The Gaussian curve (normal distribution)

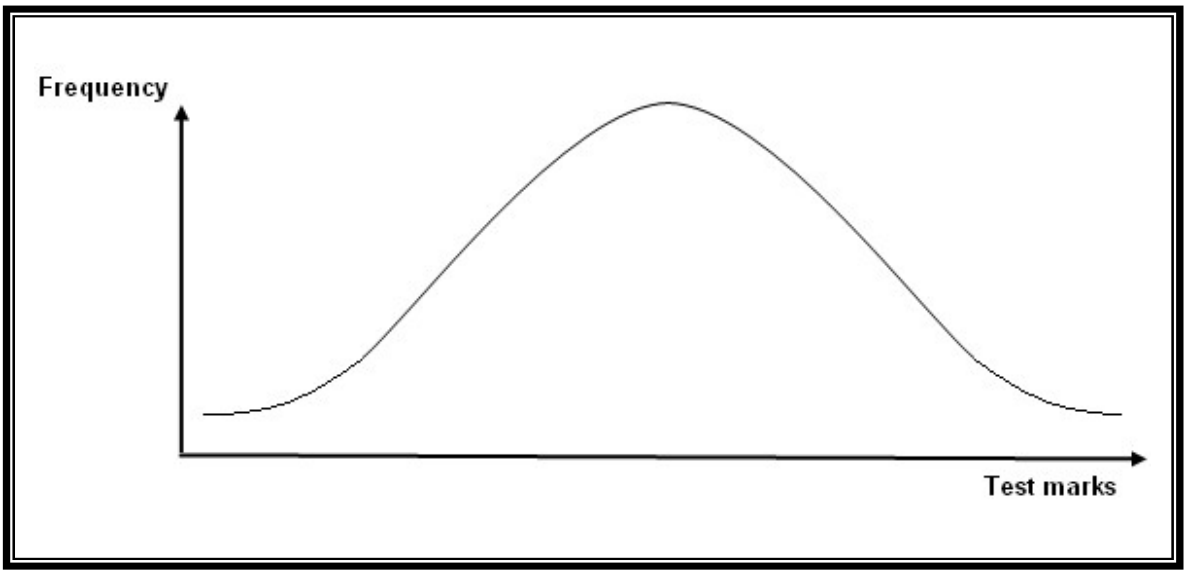
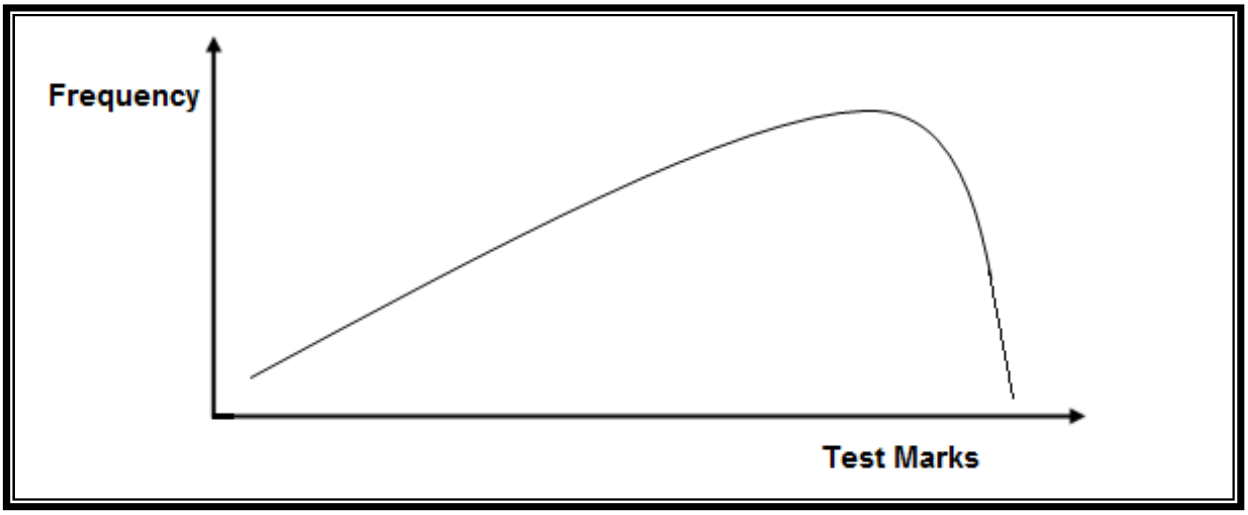


Diagram 4.2: Extreme distribution (negatively skewed)



The use of tests in schools is supposed to be a criterion-referenced type because tests are constructed according to teaching objectives. The purpose is to inform learners about their achievement and the school about the merit of different aspects of the education process. However, judging school test items within the criterion-reference concept might be possible for individual schools cases, in which an optimal teaching-learning and administration culture is available. This is rare in Saudi schools; therefore, the 'normal distribution' of test is still valid for representing the current situation of present Saudi schools' tests. My point above in reference to the interference between

the two concepts of criterion and norm referencing is a justification of making use of the first one as a framework of test items' construction process that this research project is introducing to the sample teachers and the second one as a framework of judging these items upon their function in discrimination within IAT process. The latter is represented by the way test results are interpreted where the normal distribution of difficulty or discrimination is essential to judgement. These test quality indicators will be introduced in Section 4.4 where IAT is further illustrated.

4.2.2 External Vs. Internal (teacher-based) Assessment

External assessment is applied by parties other than school or any of those related to it. Examples of external assessment in the UK are national tests such as SAT and tests that are applied by evaluating agencies such as OFSTED. Internal assessment on the other hand, is that applied by the school or by the teacher: teacher assessment (TA). Historically, assessment has been considered to be a statistical function without considering its role for effective teaching or learning process. Testing tends to be an end in itself without being incorporated instructionally (Shepard, 2000: 1). External assessment is the vehicle of this vision, which Gipps calls the traditional psychometric framework (PF) of assessment. She considers PF to be no longer adequate and calls for a paradigm shift to a more comprehensive framework that provides a "set of interrelated concepts" and opportunity to look at different aspects and dimensions of the educational practices and consequences (Gipps, 1994: 1). This is achieved mainly by internal assessment, or TA, which is a formative assessment: a framework referred to by Gipps as educational measurement framework (EMF). She discussed two assumptions that underlie the psychometric framework (PF): universalities and unidimensionality and showed their weaknesses compared to the new paradigm's concepts (EMF). She also criticised IRT theory, as a major external assessment theory, being under the assumption of one factor that underlies their prediction, whilst in reality this is not the case (Gipps, 1994: 5-7). Shepard stresses that for assessment to aid learning, content and character of assessments should be improved with the use of assessment information through TA (Shepard, 2000: 1). EMF is 'constructively' identifying pupils' strengths and weaknesses in order to assist them to improve (Gipps, 1994: 8-9). Shepard articulated this notion clearly:

> The purpose of assessment in classrooms must also be changed fundamentally so that it is used to help students learn and to improve instruction rather than being used only to rank students or to certify the end products of learning (Shepard, 2000: 31).

This research does not essentially relate to psychometrical approach and it considers that EMF is the vehicle for learning. Nevertheless, it acknowledges that some psychometric techniques are very likely work to improve TA, which in turn reflects to improving teaching practices.

4.2.3 Assessing HCD Skills

Questioning is one of the most powerful instruments for eliciting learning during instruction. According to Cotton (1988), research reported that teachers spend 35–50% of their instructional time questioning. It might be worthwhile to look at two points of view on which level of cognition questioning should take place: low order or higher order levels? Gage and Berliner (1984) demonstrated that teachers' tendency towards lower order questions is justified by the idea of "such questions bring out the raw information needed as the basis for higher mental processes." Furthermore, a study by Brophy & Evertson (1974) pointed out that for the third grade pupils, the percentage of correct answers in the class correlated positively (0.51) with achievement for lower achieving pupils. This view calls for using less difficult questions that are represented mostly by the lower order questions. Nevertheless, Redfield and Rousseau (1981) contrasted this view with fourteen experiments that resulted in a "fairly large" positive effect on pupils' achievement whose teaching included a "predominant use" of higher order questioning. The paradox revealed by these two views was explained in terms of the grade level of pupils where the former view is seen with pupils of grade 1-5, whilst the latter for those in higher grades (Gage and Berliner, 1984: 635-636). Also, research reviews by Winn (1979) and Dillon (1982) revealed that pupils' responses tend to meet similar levels of cognition (or difficulty) of their teachers' questions (ibid: 633-636). The WYMIWYG concept by Hummel & Huitt (1994) stands for: What You Measure Is What You Get. Part of their explanation of this concept is that teachers' practices of constructing HCD assessments result in their instruction and their pupils' responses to be at HCD level. Interestingly, Cotton (1988) highlighted in her systematic review of 37 studies about questioning that:

> Quite a number of research studies have found higher cognitive questions superior to lower ones, many have found the opposite, and still others have found no difference. The same is true of research examining the relationship between the cognitive level of teachers' questions and the cognitive level of students' responses. The conventional wisdom that says, "ask a higher level question, get a higher level answer," does not seem to hold.

She listed a number of findings that show this further:

1. HCD questions are not better than LCD questions in eliciting HCD responses or in promoting better learning.
2. Simply asking higher cognitive questions does not necessarily lead students to produce higher cognitive responses.
3. LCD questions are more effective than HCD questions with young (primary level) children, particularly the disadvantaged.

Nevertheless, these conclusions do not undermine the merit of questioning absolutely because, similar to what Gage and Berliner (1984) have found; reviewed research by Cotton (1988) confirmed, as the following list of findings show, that the impact of HCD questioning is seen at grades higher than primary:

1. Increasing the use of higher cognitive questions (to considerably above the 20% incidence noted in most classes) produces superior learning gains for students above the primary grades and particularly for secondary students.
2. In most classes above the primary grades, a combination of higher and lower cognitive questions is superior to exclusive use of one or the other.
3. For older students, increases in the use of higher cognitive questions (to 50% or more) are positively related to increases in [number of aspects; among these are]:
    (a) The number of relevant contributions volunteered by students
    (b) The number of student-to-student interactions
    (c) Speculative thinking on the part of students
    (d) Relevant questions posed by students

Acknowledging this notion, this research design considers that practicing HCD questioning during instruction is very likely an optimal factor for promoting HCD critical thinking by pupils of the higher grades' levels: intermediate and secondary schools that this research targets.

4.3 The Role of Teacher-based Assessment in Saudi Arabia

For TA, and from my observations and long experience of working with Saudi teachers and schools, Saudi teachers that give attention to assessment as a major component of their instructional role consider its function as a motivating factor in how well pupils study. Nevertheless, they rarely consider this assessment function as a way

of providing feedback for their professional development, either for instruction or for design testing. Another group of teachers tend to consider assessment as a proxy to certification and as one of the many "job duties" that they have to undertake. Unfortunately, the second group is the majority and many schools do not give such an issue much consideration as soon as pupils succeed and "the job is done." There are no criteria based on TA quality that schools can be evaluated upon. Consequently, this sort of practice has spread over the past years and both teachers and schools have begun to consider it the "default" setting for teacher-based testing, i.e. "get the job done." Some teachers have developed a further technique to overcome the problem of criticism when somebody attacks their "overly easy" exams. They prepare pupils for the test, which is accomplished through worksheets that include questions of various difficulty levels but at the same time represent most of the test's questions. This practice, together with the fact that there are no standardised national tests that schools and educational authorities can rely on, forms the bipolar dilemma of the TA role in Saudi education. Moreover, an appraisal of teachers' testing practices is not on any agenda of the educational supervision authorities, as will be elaborated on next.

Firstly, teacher testing ability is not included in the *employee performance evaluation form* that is used annually by the school principal to score teacher performance. This form was improved four years ago, which is reflected in Appendix 1 showing the previous and the recent forms. However, I draw on the former, since this discussion presents that form's historical impact. Item 7 only deals with this aspect by a general statement lacking accuracy. Al-Dakheel (1997) investigated the efficacy of this *employee performance evaluation form* from the perspective of Riyadh school principals and educational supervisors. Among what she found was a lack of following-up on performance evaluation results. She also found that approximately 68% of the respondents believed that this form contributed in diagnosing teacher weaknesses, which leaves almost 32% who did not share the same opinion. I believe that this level, though not so high, needs to be considered when looking at the critical role of this form's contribution to the quality of education. The study have also found that the percentages in areas of estimating the competency of teachers, achieving educational aims, and revealing training needs, are in a similar range.

Secondly, when educational supervisors do look at teachers' tests, this type of review has two main shortcomings: (a) it is not a scheduled or compulsory task for each supervisor, so it is done upon the supervisor's initiative and (b) it is done only for the

questions sheet, without looking at pupils' specimen answer-sheets and without giving each teacher a separate report but provide a general report for all observed pitfalls. Training teachers to write better test questions has been of the educational supervisor's concern recently, which might be a response from them to improve their current limited efforts.

Research about *Diploma Disease*, or *'paper-qualifications syndrome,'* has indicated the effect of social role selection for employment that results in minimising educational aims. Ronald Dore and John Oxenham, as two leaders of the thesis of *Diploma Disease,* defined the interactions between employment and educational qualification. Oxenham questioned why employers use schooling as a basis of qualifications for jobs (Oxenham, 1984). The thesis cited examples from England, Japan, Sri Lanka, and Kenya. Furthermore, empirical studies have focused on the dysfunctional effects of this phenomenon by examining how Sri Lanka, Mexico, Ghana, Tanzania, Kenya, Somalia, Zambia, Sierra Leone, Liberia, and the Gambia use educational qualifications for job recruitment only. Also, the impact of this phenomenon on teaching and learning quality was explored in Ghana, Mexico, China, Malaysia, Chile, India, Iran, and Thailand. Various levels of analysis were targeted when studying this issue, including the recruitment and selection systems of modern sector labour markets, the educational ethos of schools, learning processes, and labour force learning and motivational orientation (Little, 1993).

Furthermore, Keith Lewin (1985) explored the relationship between selection and curriculum, and highlighted the reasons for the reported effects on curricula as a result. Among these was the assessment emphasis on LCD skills, embodied primarily in the "recall of trivial information." The resolving strategies to be tackled, he recommended, should consider different aspects, some relating to pupils' and parents' motives and others to assessment policy and practices (Lewin, 1985). Moreover, Nigel Brooke and Oxenham explored the influence of certification and selection on teaching and learning. Angela Little examined evidence on whether schooling is associated with productivity in a variety of jobs. She also evaluated the possible effects of proposed solutions to *Diploma Disease* (Oxenham, 1984).

Although the situation in the KSA is different, it has some relevant implications for *Diploma Disease* that can be seen in a number of aspects. These include the tendency towards using final examinations as a promoting device to a following level, the intensive use of LCD, the absence of standardised assessment, teachers' practice of

preparing pupils for exams, and the lack of well-structured selection procedures for employment, especially given that certificates are unquestioned factors for judgement.

4.4 Item Analysis Technique (IAT)

4.4.1 IAT under the framework of Classical Test Theory (CTT)

IAT is a technical procedure that produces two coefficients that indicate the level of strength or weakness of a testing item. These two are namely: difficulty coefficient (*P*) and discrimination index (*D*). Their names reveal their function where *P* is used to indicate the extent to which an item was difficult or easy: the more the *P* value the easier the test is, hence sometimes *P* is called 'easiness' index or 'facility' index; nevertheless, in literature of testing it is more common to call *P* as difficulty index therefore this term will be used here. The formula that calculates it for questions of dichotomous scores (which result from closed ended questions) is:

$$P = \frac{\textit{No. of succeeded students in that item}}{\textit{No. of all students that attempted to answer}} \%$$

The formula that calculates it for questions of polytomous scores (which result from open-ended questions) is:

$$P = \frac{\textit{Total of students' scores for that item}}{\textit{No. of students} \times \textit{Standard score of that item}} \%$$

Following the norm referencing approach[24], the optimum value of *P* is that around 50% (or 0.5) and accepted values range from 0.3 to 0.9; however, there are some considerations in terms of the type of the item under judgement where those of two alternatives differ from those of 3 alternatives and so on. Kearns (not-dated) mentioned that recommended *P* value for multiple-choice questions of 4 alternatives is 63% up and for true-false question is 75% up. These results are calculated from the following formula:

$P = 0.5 + 0.5\ (1/Alt)$ where Alt = number of alternatives

The other parameter, discrimination index (*D*), also called Findley's *D*[25], is used to clarify how much that item was able to discriminate between high performing students and low ones. This comes from the fact that an ideal question is not supposed

to be answered evenly amongst all students; if so, then there is no meaning of asking such question. It is assumed that an ideal question should reveal the set of students who are those most able and who are those less able (Osterlind 1992: 283). Statistically, it is considered that *D* "indicates the degree to which responses to one item are related to responses to the other items in the test" (Allen and Yen 1979: 120) The formula of *D* calculation for questions of dichotomous scores is:

$$D = \frac{U - L}{n} \%$$

where,

U: number of students from the upper half group that answered this item.

L: number of students from the lower half group that answered this item.

n: number of upper or lower group (equal).

The formula that calculates it for questions of polytomous scores (which result from open-ended questions) is:

$$D = \frac{U^{scores} - L^{scores}}{n \times St^{score}} \%$$

where,

U^{scores}: total of scores obtained by all students of the upper half group.

L^{scores}: total of scores obtained by all students of the lower half group.

n: number of upper or lower group (equal).

St^{score}: standard score of this item.

Following the norm referencing approach, the higher the *D* value of an item the better the quality of that item. Ebel (1965b: 364 as sited in Stanley and Hopkins 1978: 273) highlighted the "rules of thumb" of *D* values as shown in Table 4.1 below.

Table 4.1: Ebel's roles of thumb for *D* values

D	***Judgment to the item.***
0.40 and up	Very good item.
0.30 - 0.39	Reasonably good item, but possibly subject to improvement.
0.20 – 0.29	Marginal item, usually needing and being subject to improvement.
Below 0.19	Poor item, to be rejected or improved by revision.

There are different formulas for calculating *P* or *D,* but the aforementioned have been presented for the purpose of illustrating their concepts through their mathematical meanings. For further detail about calculating *P* or *D* by different algorithms and for

objective or subjective question types, the reader could find such detail in Davis (1951), Wood (1960), Ebel (1965) and Brown (1970) (as sited in Brown 1970, 277) and in Stanley and Hopkins (1978: 267-280). It is important to clarify that optimum *P* and *D* values, as explained above, are not to be taken solely, but with concurrence of each to the other in terms of the purpose of the test.

Another indicator that should be taken into account while judging a test's item is power of distractors. Distractors are those alternatives in multiple-choice questions other than the correct one. Distractors that no students have selected or distractors that 'fool' all students, high and low students alike, are considered poor quality. The distractor that the majority of students choose is likely to be a correct answer (Kehoe, 1997). I have to mention that there are other coefficients that give an indicator to reliability or internal consistency of a test such as Cronbach Coefficient: Alpha, Kuder Richardson (KR) Formula, or Spilt-half Reliability Coefficient (Yu, 2002). However, discussion of these indicators is beyond the scope and aims of this research because these parameters inform the quality of the whole test whereas the present research's focus is on each test item to enlighten the teacher about her/his related ability.

4.4.2 IAT under the framework of Item Response Theory (IRT)

IRT is a modern statistical framework for measuring quality of test items. Compared to Classical Test Theory (CTT), it is free from test characteristics such as size, administration, speed and the sample of students upon which it has been standardised. However, it is dependent on item's characteristics only. In designing a research for analysing test questions, one should decide which of the two frameworks to utilise. Recently, IRT is acknowledged as a sound scientific and mathematically robust framework for analysing tests, but still, there are many preferring CTT methodology for a reason or another. In the following presentation, I need first to explain both frameworks' main concepts and methods then to present and justify my position in this respect.

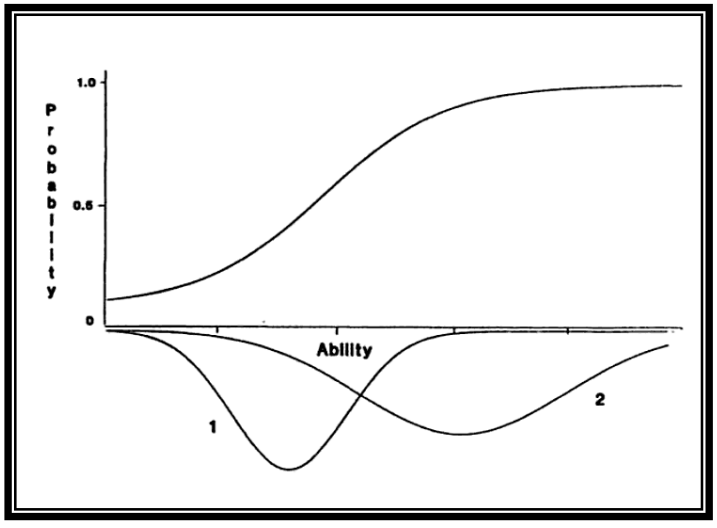
The item statistics of interest of IRT three-parameter model are: b, which stands for item difficulty; a, which stands for item discrimination; and c, which stands for pseudo guessing parameter. (for more see for example Lord, 1980, Hambleton & Swaminathan, 1985, Hambleton, Swaminathan & Rogers, 1991). Its major concept is that performance of any student on a test item is to be explained by a set of factors called latent traits or abilities (Stage, 1998:3-4) -thus it is also called latent trait theory.

CTT on the contrary is dependent on the specific sample of students for the pilot test and hence item analysis results do not necessarily inform a later test administration.

It is thought that CTT has its limitations to solve some testing problems such as design of tests, identification of biased items, adaptive testing and equating of test scores. IRT plays better role in this respect and provides a model that is expressed at the item level rather than the test level because item characteristics described by IRT are not group-dependent (Hambleton et al., 1991: 5)

There are many models of IRT according to the number of parameters to be found. These parameters are to describe the item or the examinee. However, an IRT model may or may not be appropriate to explain a particular test data unless "the fit" of that model to the studied data is assessed (Hambleton et al., 1991: 7). The popular models of IRT are one-, two-, and three-parameter logistic models, corresponding to the number of item parameters associated with each. Item characteristic curve (ICC) is a major method of IRT that expresses mathematically the probability of success on an item in relation to the ability measured by the test and the characteristics of the item (Hambleton et al., 1991: 12). Diagram 4.3 shows an example of ICCs.

Diagram 4.3: An Item Characteristic Curve (ICC) and distribution of ability for two groups of examinees (Hambleton et al., 1991: 8).



In Table 4.2, a comparison summary is provided between the two approaches CTT and IRT. However, there is little interest in research to compare CTT and IRT in terms of item analysis and test design since the focus is on test equating primarily.

> It is somewhat surprising that empirical studies examining and/or comparing the invariance characteristics of item statistics from the two measurement frameworks are so scarce. It appears that the superiority of IRT over CTT in this regard has been taken for granted in the measurement community, and no empirical scrutiny has been deemed necessary. The empirical silence on this issue seems to be an anomaly. (Fan, 1998 p. 361 as sited in Stage, 1998: 4)

Table 4.2: Comparison summary between IRT and CTT approaches.

Aspect of Comparison	IRT	CTT
Purpose	Measure quality of test items	Measure quality of test items
Level	Provides a model that is expressed at the item level rather than the test level.	Describes the item in the context of the test only.
Dependence	Free from test characteristics such as size, administration, speed and the sample of students. Depends on items' characteristics only.	Depend on test characteristics such as size, administration, speed and the sample of students.
Relevance	Item analysis results inform any later test administration.	Item analysis results inform later test administration that are similar to the pilot one.
Concepts	Considers the concept of 'latent' traits of the individual.	Does not consider latent traits.
Problem solving	Ideal for solving some testing problems such as design of tests, identification of biased items, adaptive testing and equating of test scores.	Limited in solving these testing problems.
Size of sample	Requires that the sample should be of a big size otherwise it may not give considerable findings.	Works with big size as well as small size of samples such as one classroom's students.
Mathematical concepts and calculations	Uses advanced mathematical concepts and their calculations are difficult and time consuming.	Uses common mathematical concepts and their calculations are attainable but time consuming..

Stage (1998) has studied data of the Swedish Scholastic Aptitude Test (SweSAT) by the two theoretical frameworks and found that the difference is minimal considering that "it does not seem to be of any importance for the test design whether the item analysis has been performed within the IRT framework or within the CTT framework" (Stage, 1998: 33). It should be noted that IRT popular models are appropriate for dichotomous items only. Furthermore, IRT requires that the sample

should be of significant size otherwise it may not give considerable findings. For classroom size purposes, as the case of teacher assessment in this work, it is enough to use CTT. Antelmo et al. (2005) explained that despite IRT is the pre-dominant measurement model, CTT is still frequently employed because of many reasons. First, a big sample size, far greater than classroom size, is necessary for IRT application. Second, many popular software programs, such as Excel, are utilisable to calculate CTT indicators while IRT needs specialised packages such as Bilog, Winsteps, Multilog, RUMM. And third, the simplicity of CTT techniques compared to IRT's. They believe that "tutors without a strong statistical background could easily interpret the [CTT] results without going through a steep learning curve." These reveal justifications for the present research to focus on IAT under CTT only, since it represents a framework that is sufficient for measuring TA. In fact, the underpinning calculations for CTT come from common mathematical concepts, ratio and percentage; hence this could effectively provide teachers with a minds-on experience for the new intervention. Moreover, reading outcomes of IAT under CTT is easier than reading IRT outcomes. For example, within IRT, the teacher needs to be familiar with reading curves, especially comparing many ICC curves that come together on one schema as Diagram 4.3 has shown. Some other required skills for reading IRT findings might be more difficult for majority of teachers.

4.5 Computer Aided Item Analysis Technique (The CAIAT)

A number of software packages perform test item analysis under CTT or IRT. In Section2 of Appendix 1, I present a systematic review for two common CTT solutions, namely ITEMAN© and Lertap©. The review shows why the CAIAT represents a better alternative for this research's purpose by highlighting its major characteristics, which are as follows:

1. Works with an Arabic language interface.
2. Uses driven menus and dialogue box forms for interacting with the user, and is thus characterised by its ease of use and ergonomic advantage. In fact, every user who exhibits basic computer skills and understands IAT concepts can use it with minimal training, or even without any at all.
3. The user enters raw data relating to test item marks by different means.
4. Outputs can be seen on-screen or printed out in tabulated forms that are clear, simple, and understandable to teachers, principals, and supervisors.

5. Resulting analysis indicators can be calculated for both types of questions: objective and essay questions.
6. A special processing of statistical calculations is related to multiple-choice items by calculating the distraction efficiency index for every alternative.
7. It has the facility to store the entered data on databases that can be retrieved later, so it can be used on one PC by many teachers, who can store his or her data on an independent file, without interference from others.

As an advanced programmer using MS Visual Basic©, I have developed this software package in its entirety[26]. I have designed its structure, screens, and reports and have written the program code responsible for calculating item analysis parameters *P* and *D*. I have also followed up its application and made any necessary improvements or developments to improve performance to an optimal level that meets the requirements of facility and ease of use that this research encourages.

Chapter 5
Method and Design

5.1 Method

5.1.1 Evaluative Research

The purpose of this research is to evaluate the feasibility of the HCD/CAIAT project as a package that can be used to promote teachers' PD in HCD question construction skills. Evaluative research is considered one of the most modern approaches to social research, connected as it is to the emergence of social welfare (Soydan, 1998). However, many commentators have indicated the political dimension of evaluation, since most evaluation projects are in response to decision makers' requests. According to MacDonald (cited in Cohen et al., 2000, p. 40), "evaluation is an inherently political enterprise." One of the comprehensive forms of the many definitions applied to evaluation is "the systematic acquisition and assessment of information to provide useful feedback about some object" (Trochim, 2006). Although many differences have been highlighted between evaluation and research, Norris (1990) suggests that they share similar methodologies, and as such the former is an extension of the latter (Cohen et al., 2000, pp. 38-42). In this respect, *evaluative research* is appreciated as a moderate form that encompasses the benefits of both approaches. However, it is deemed as theory-driven rather than theory-generating – as found in pure research (ibid., 2000, p. 38). With good attention to the considerations of bias and the control of variables that might threaten the role of evaluative research, I believe it is very likely to reach higher levels of reliability that places it on the same level of credibility as pure research, especially when one considers the absence of stakeholders involved in funding, monitoring, intervening in, or judging the work. In the present project, no such political aspects are imposed on its existence or purpose; therefore, it is much more within the category of pure research than that of evaluative reporting. For this reason, in the following sections I am going to present methodological considerations pertaining to research.

5.1.2 Quantitative Or Qualitative Approach?

Quantitative research uses methods that are designed to ensure objectivity, generalisability, and reliability; this is by means of statistical methods used to test

predetermined hypotheses regarding the relationships between specific variables (Palya, 2000). Its advantage is that it measures the reactions of a great number of individuals, thus facilitating comparison and statistical treatments of collected data. This aids the possibility of generalisation. It is said that "Quantitative methods offer a higher degree of precision, consistency and reliability of findings" (CERIS, 2004: 23). The weakness is that it treats human behaviour in a way that removes the event from its context. When used alongside qualitative method, this shortcoming is mostly overcome by a meticulous awareness of the researcher.

Qualitative research method is designed to provide a perspective of targeted audience members through immersion in a culture or situation, direct interaction with the study's individuals, and to gain a holistic understanding of the studied problem (CERIS, 2004: 23). Qualitative methods include observations, in-depth interviews, and focus groups. These methods are designed to help researchers understand the meanings people assign to social phenomena and to explain the mental processes underlying behaviours. Hypotheses are generated during data collection and analysis, and measurement tends to be subjective. In the qualitative paradigm, the researcher becomes the instrument of data collection, and results may vary greatly depending upon who conducts the research. This could colour the results of the data collector's views or backgrounds, which calls to categorise qualitative research's results as relative rather than absolute. Nevertheless, the advantage of using qualitative methods is that they, through optimal implementation, generate rich and detailed data that leave the participants' perspectives intact. A practical difficulty however, is that data collection and analysis may be labour intensive and time consuming.

There are considerable writings about the integration of both methods, each of which is either explaining theoretical considerations of such a trend or presenting practical guidelines for the proper adaptation of one with the other (e.g. Bryman 1988; Brannen 1992; Cresswell 1994; Erzberger 1998; Erzberger & Prein 1997; Denzin 1978; Flick 1992; 1998; Fielding & Fielding 1986; Kelle & Erzberger 1999; Tashakkori & Teddlie 1998, As cited in Kelle, 2001). One of the present research's goals is to encourage decision makers to adopt the HCD/CAIAT project as an instrumental method for improving education. Some types of information needed for decision-making require quantitative data, such as detecting any measurable differences in knowledge and/or attitudes. Other types are preferred to be of qualitative data, such as investigating

individuals' reactions, or practical skills. To be effective and responsive, this evaluative study should rely upon a combination of the two approaches of data collection, with an emphasis on the latter to learn as much about the target audience as possible. More notably, the qualitative approach can provide the sort of data generated by the quantitative approach; therefore, some qualitative methods in this research (such as case studies) will provide measures of acquired knowledge by the participants. This outlines the importance of the qualitative approach followed in this study. The most important point about triangulation is that shortcomings of quantitative method could be overcome by qualitative method and vice versa.

This research will follow the descriptive method design for answering its research questions. In educational studies, it is common to use hypotheses in a quasi-experimental design. Although this study is not a quasi-experimental design, I have used hypotheses for the purpose of organising some of the research questions' inclusion into a form that serves the applied statistical treatments. Ary, Jacobs, and Razavieh (2004, p. 107) indicated that research questions cannot be examined directly by statistical analysis and need to be rephrased in hypotheses forms. They stressed that research hypotheses "guide the process of data collection and interpretation and assist the researcher in identifying what path should be followed and what kind of data should be collected." Some other authors confirmed this further that when a research study sets out to explain facts or phenomena, the use of hypotheses becomes essential (Obaidat, Adas, and Abdulhaq, 1996).

There are two types of hypotheses: the *alternative hypothesis* and *null hypothesis* (Somekh & Lewin, 2004). The first is phrased in a determining statement that identifies the magnitude and direction of the relationship between studied variables, or in other words it indicates that there are differences. As a result of this level of determination, this type of hypothesis is used when the researcher holds evidence about claims underpinning the hypothesis. This evidence could be previous studies, a minor study that precedes the major study, common sense, etc. (Dalen, 1962). The second type of hypothesis, the null hypothesis, is phrased in a null statement that indicates that there are no differences. This type is also called the "hypothesis of no difference" (Kumar, 2005) and is used when the researcher has little or no sufficient evidence about the studied phenomenon. Actually, this applies to most of research situations; therefore, it is used frequently in educational research (Badr, 1989). In the present research, I need

to use the null hypothesis because, as reviewed previous studies have shown, there is a paucity of previous related work available to support generating a non-null hypothesis.

5.2 The Project Design

5.2.1 Sampling in Educational Research

Sampling is conducted in educational research to provide researchers with a way of measuring phenomena, without the need to approach the whole population. Statistical samples are designed to provide the same results obtained when the whole population is treated. This notion entails that such findings be generalised to the whole population, which is the ultimate goal of most social and educational research. Ross (2005) mentioned a number of advantages of sampling compared to studying the whole population, especially when the latter is of a huge size. Among these are reduced costs and improved speed in data summarisation and reporting. He mentioned three categories of research types for sampling, each of which requires a specific treatment for selection and validity considerations. These are experiments, surveys, and investigations, the latter two of which are employed in the present research. McMillan (1996) believes that the degree of sample representativeness depends on the applied sampling technique.

The accuracy of findings generalisation is derived from statistical theory which requires that a probability sample type is used. Probability sampling occurs when "each member of the defined target population has a known, and non-zero, chance of being selected into the sample" (Ross, 2005). Commonly used examples of probability sampling are simple random sampling, stratified sampling, and cluster sampling. Usually, it is considered in research methodology that probability sampling is the best for ensuring research quality, and errors included in this type can be minimised to an acceptable level by making the sample size, statistically, large enough (SurveyMethods, 2009). Specific statistical tables aid researchers for this purpose. Probability sampling is the prime method used by quantitative research, although it is also used by qualitative research. Non-probability sampling is the other alternative used frequently by qualitative research.

According to the SurveyMethods (2009) website, probability sampling is considered stronger than non-probability sampling and provides more accurate data that fit for generalisability. This concept is widespread throughout the research literature. Nevertheless, qualitative researchers believe that non-probability sampling could be

designed as valid for generalisation purposes. In this respect, Maxwell (2007) highlighted generalisability to theory rather than to the population, citing Robert Yin's (2003) notion of "analytic generalisation":

> Analytic generalization is not generalization to some defined population that has been sampled, but to a theory of the phenomenon being studied, a theory that may have much wider applicability than the particular case studied. In this, it resembles experiments in the physical sciences, which make no claim to statistical representativeness (physicists don't draw random samples of atoms), but instead assume that their results contribute to a general theory of the phenomenon.

Gobo (2007) pointed out that this type of generalisation concerns the "nature of a process" and is based on the notion of "theoretical sampling," which is defined by Glaser & Strauss (1967) as follows:

> Theoretical sampling is the process of data collection for generating theory whereby the analyst jointly collects, codes and analyses the data and decides what data to collect next and where to find them, in order to develop the theory as it emerges (Glaser & Strauss, 1967).

This is part of the *Grounded Theory* approach initiated by Glaser & Strauss; however, the sort of theoretical sampling meant by Gobo is not practically identical to that utilised for grounded theory but instead borrows the concept. To elaborate, Gobo (2007) argued that sampling units are different from observational units, and statistical population is different from social population. The latter is also described by the terms "logical" versus "social universes," meaning that the researcher should clearly distinguish between these two universes while sampling. In identifying which variables need to be taken into account for social sampling, instead of misleadingly confined to the usual statistical sampling practice he highlighted the concept of the variance of the sampling unit (or the variance of the phenomenon under study). Identifying the unit under study is the first crucial step. For example, in Gobo's research studying "customer relationship management" for call centres, instead of clustering the sample into the usual statistical organisational categories of private, public, and non-profit centres, he derived the clustering from the variance of the phenomenon under study (customer relationship management), and in this sense, customer relationship is considered critically dealing with counselling, marketing, interviewing, and advertising. The other set of classical variables seem to have little association with studying how customer relationship is managed compared to these. That is to say that instead of studying the statistical differences for customer relationship among the private, public,

and non-profit centres' customers, it is better to study those differences among the customers that are subjects to counselling, marketing, interviewing, and advertising because these variables are the ones that underpin the variance of the phenomenon under study (or sampling unit). The second step in social sampling is to plan for representativeness upon the level of variance of the study unit. The more the variance, the more the sample size, including all types of cases that cover that variance (Gobo, 2007). For the example above, when choosing a 'typical' sample that includes customers from all the four areas of counselling, marketing, interviewing, and advertising there might be none of those customers from, say, private centres then this would not affect representativeness since private centres do not related to the variance of the chosen sampling unit in this case. The size and demographic characteristics of the population are not the prime determinants of this sampling, as in statistical sampling, although they might be included according to their relevance to the phenomenon under study.

5.2.2 Research Population and Sampling Method

In his book written for the International Institute for Educational Planning, UNESCO, Ross (2005, p. 7) explained the characteristics of an optimal *judgement* sample:

> The process of judgement, or purposive, sampling is based on the assumption that the researcher is able to select elements which represent a 'typical sample' from the appropriate target population. The quality of samples selected by using this approach depends on the accuracy of subjective interpretations of what constitutes a typical sample.

Considering this point, I have selected in the present research to use this type of non-probability sampling, i.e. judgement (or purposive) sampling, adopting at the same time the concept of 'typical sample' as I will explain shortly. The population of the study includes all female science teachers across the KSA, and I consider Al-Ahsa female science teachers as a 'typical' cohort, particularly given that there are many shared characteristics among all teachers of the subject. In the KSA, there are similarities in many aspects of education provision in different provinces and cities of the state; among these are policies on education, the curriculum, systems of employment, governing rules and educational offices, along with teachers who come from Saudi universities that again share similar aspects of higher education provision. Actually, these aspects are the results of the central systematic control of the state's governing policy; as such, I

will take this point for, purposively, considering Al-Ahsa female science teachers as a 'typical sample' representing the whole of the KSA's female science teacher population.

On the other hand, in utilising Gobo's notion of variance of the phenomenon studied, I have identified my 'sampling unit' as the *effectiveness of the CAIAT in improving science teachers' ability in test item construction*. The theory of the research has little relationship to the demographics of the ultimately targeted population of KSA teachers. The variables of interest here relate to characteristics that portray teachers' knowledge and abilities rather than where geographically do they teach. Therefore, the Al-Ahsa sample fits as a 'typical' sample that includes most of the variables required for examination by the research theory in order to assess its assumptions. These variables are: educational qualification, prior training on test construction skills, prior training on IAT, key stage (intermediate/secondary), level of graduation, years of experience, and specialisation subject (physics – chemistry – biology). Educational research in the KSA usually considers such variables rather than geographical location because of the centralised nature of the Saudi educational system. Actually, I have not come across one local research in education that study examining the differences between different regions' teachers. Therefore, the generalisability of the Al-Ahsa findings to the population is possible according to Gobo's conceptual argument of social research. Furthermore, keeping in mind what Robert Yin's (2003) notion includes about generalisability to theory rather than to population (which he called analytic generalization), and the fact that this research is attempting to examine a theoretical assumption of the CAIAT functionality, justifies this research's compatibility with 'typical' sampling.

It is essential to remember that education in the KSA is provided in single-sex schools whereby boys' schools have male teachers only and no females are allowed to mix in these schools. Girl's schools have female teachers only and no males are allowed to mix within these schools. Consequently, the application of this research project could be undertaken in one of the two sectors of education. In the following two sections, I will explain which sector was selected for each situation and why.

5.2.3 Pilot Sample

The first phase involved a pilot study to measure the validity and reliability of the data collection instruments and aided in revealing points of consideration that I should take into account for the following application on a greater scale for the main

sample. The pilot study's sample was chosen as male teachers because I could participate personally and gain hands-on experience in the administration of the data collection instruments and the different activities of the fieldwork such as the workshops, presentations, technical support for the CAIAT software, etc. In the second phase, I cannot have such personal involvement, since this was carried out with female teachers, with whom I cannot mix according to the policy of single-sex schooling in the KSA. The pilot application was used as means for the researcher to update research instruments. Among these was the CAIAT software package, completely developed by the researcher, so the sort of errors or bugs that could occur during the very first implementation needed my first-hand supervision as the software programmer. The sampling of males and females separately should not be looked at as an intention to study gender differences, but is only the result of sampling within a single-sex schooling context according to the cultural/social considerations mentioned above.

I confined the pilot sample to one specialisation (physics) and a smaller number of individuals (42) to gain better knowledge for myself as a researcher on how this project would work at the first attempt. Having fewer people to work with in the training, in the CAIAT installation, in the program running, in the monitoring, in discussions and following-up gave me a better opportunity to concentrate, analyse, and discover what may have not been apparent to me during planning. This informed the second phase involving the main sample very richly. I redesigned the training time and sessions, reprogrammed some of the CAIAT functions, and updated some of the data collection instruments as an outcome of the pilot application.

The pilot sample selection was drawn from figures from 2004, when the high schools population of male teachers in the KSA amounted to 30,764 (MoE website, 2005). In Al-Ahsa, the figure was 1,363 teachers for the same year and there were 56 boys' high schools (MoE website, 2005) with 87 physics teachers (GDBEA, 2005). Out of these, I selected a sample of 50 teachers, mainly from the city of Al-Ahsa and excluding rural school teachers because of difficulties accessing some of these areas[27]. The same consideration applied to selecting the main sample's teachers. However, not all of the 50 pilot sample's teachers that were selected and invited attended the project activities, and so only 42 were able to attend.

5.2.4 Main Sample

This sample was applied to girls' schools, meaning that the teachers were female only. I chose this sector (girls' education) in order to provide the research with a very important source of quality, which is management support. One of the main assumptions of this project is that its success is dependent on the level of management support that should be given by the local educational authority. My position as a head of girls' education facilitated, but did not dominate, the ease of application during the fieldwork with the sample of female teachers. This is represented by the provision of enough team members to apply the project and the provision of training facilities, work time and resources that the working team needed during workshops or other activities. Decision-makers in the MoE should be encouraged to adopt management support as a prerequisite of the project; hence, the final presentations of this project to stakeholders should put forward this consideration.

This sample covered two key stages: intermediate and high school[28], to account for different science specialisations. The size of this sample was 409, which is almost identical to the whole science teacher population in urban Al-Ahsa girls' schools. Although this was a non-probability sample, I will highlight one statistical strength. The selection of this sample's teachers was drawn from figures for the academic year 2004/2005, when the population of female teachers (not only science teachers but all) stood at 228,622 (MoE website, 2005). According to the statistical table used for identifying the size of the sample that corresponds to a specific size of population, the Saudi teacher population lies between 100,000 and a million and thus requires a sample size of at least 384 in order to gain a confidence level of at least 95% (see this table at Box 4.2 of Cohen et al., 2000, p. 95). The present project's sample size fulfils this because it was 409 which consolidates, from a statistical point of view, that Al-Ahsa teachers are representatives of the rest of the KSA's science teachers.

5.2.5 Case Study Sample

Another kind of sample is that of the case study (CS). Since Saudi teachers are not familiar with case studies, it was difficult to find anyone who would agree to be the subject of one. Therefore, some difficulties led to the following gradation in this sampling process. Firstly, I chose to follow the *convenience* sample type by considering those who were available to me as the researcher. This non-probability sampling

method subscribes mainly to the qualitative research approach (Fink, 2002, p. 23). In this respect, I chose 18 teachers who agreed to be part of the CS treatment. Unfortunately, many of them were reluctant and hence all did not agree to participate. Moreover, some others were absent from school for a long time vis-à-vis sick leave, maternity leave, etc. I then tried again to find other teachers who would agree to participate, but this time decided to plan to meet the *purposeful* sample's characteristic of representativeness. As I anticipated that this new round of sampling might end up with very few people, I was keen to promote diversity within this group. I selected five cases which, as Table 5.1 shows, were chosen from different subject areas, different key stages, different years of experience, and different schools.

Table 5.1: Major demographic data for the case studies' individuals

Case	Key stage	Specialization	Years of Experience
First	Intermediate	chemistry	13
Second	Intermediate	physics	6
Third	secondary	physics	6
Fourth	Intermediate	chemistry	13
Fifth	Secondary	Biology	6

5.2.6 Work Plan Scheme

HCD/CAIAT project is implemented throughout five stages: introduction, training, contingent application, mandatory application and complementary procedures. The following sections (5.2.6.1 to 5.2.6.5) explain these stages in detail, and Figure 5.1 and Table 5.2 illustrate their sequence and main functions. There is a team of research assistants which is comprised mainly of science educational supervisors at General Directorate of Girls Education at Al-Ahsa (GDGEA); and from GDBEA for the pilot sample. There are also some senior specialists from the Department of Educational Measurement and Evaluation (DEME) at GDGEA who contributed in some data collection processes. This is to increase the number of assistants, since the number of educational supervisors is not enough to cover the high number of participating teachers. However, all were trained before the fieldwork on what is required. Their role is to participate in the training course, observe teachers during workshops and lesson observation, and act as judges for the teachers' made tests that will be under this study's

investigation for quality of question construction. Other participants are from the Computer and Informatics Centre at GDGEA who took the role of technical support. My role as a researcher is training the work team, supervising their training for teachers with some intervention whenever needed. Moreover, for the pilot sample, I facilitated and observed during workshops, as well as have carried out interviews. With female teachers, I was not physically a part of the application or the field administration of any instrument, and have handed over this task completely to my major assistant who is the head of the DEME. Furthermore, I assigned her as a leader for the researcher's female assistants' team since she is qualified in doing research through her Master's degree and her on-going design, administration, and analysis of program evaluation instruments that her unit is dealing with.

Appendices 2 to 11 show the different measurement instruments and instruction documents that were used throughout the work. They can provide the reader with a good perception on how this project was implemented, so it is advised to be referred to while reading the next sections whenever needed.

Figure 5.1: A flow chart for research procedures

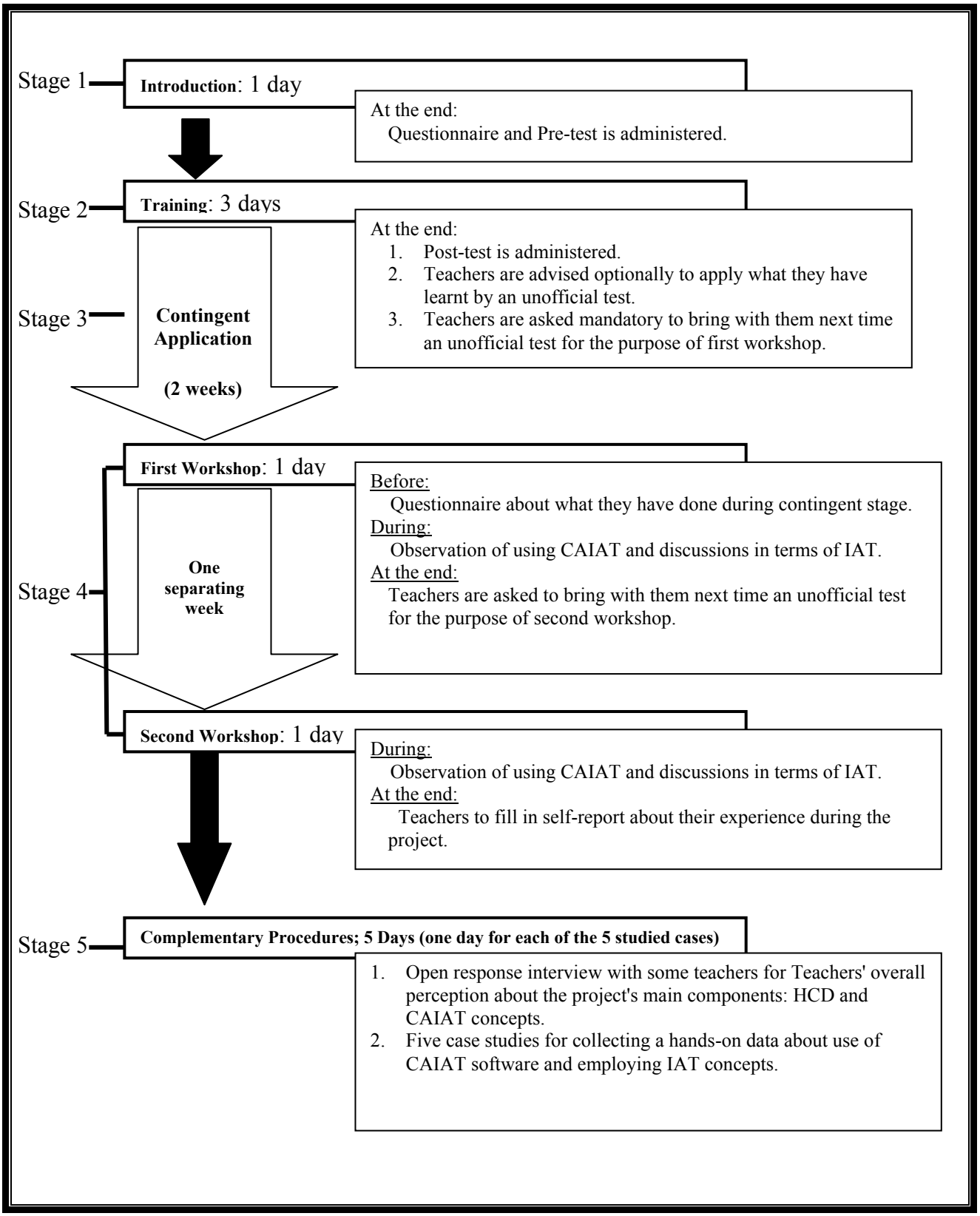


Table 5.2: Research stages' sequence and lists of corresponding procedures

Stage	Purpose	Collected Data	Instruments
1.Introduction	1.Presenting HCD/CAIAT project to the teachers.	None	None
	Measuring the pre-test results.	1.Basic data of participants. 2.background in computers, HCD and IAT.	Questionnaire/pre-test (Appendix 2).
2.Training	Training participants on three issues: HCD concept, IAT, and the CAIAT software	Non	Appendix 3 shows the training course syllabus
	Post-test: it is identical to the Pre-test that has been administered in the introduction stage.	1.Outcome of the training course in terms of application. 2. Overall efficacy of the training course.	Post Test (Appendix 4)
3.Contingent Application (optional)	At the end of training, the teachers are advised to apply concepts and skills of the project individually at home. The can use some of their own previous tests so as this could aid to show strengths or weaknesses of those tests that they are familiar with.	Teachers self-reporting about their practice.	Questionnaire of Contingent Application Stage (Appendix 5).
4.Mandatory Application and workshops			
First Application and Workshop	First application test: The teachers are asked to write unofficial test, i.e. the one that does not count for the final mark (quizzes for example); the test to be for a specific topic with concentration on HCD items and then to administer and mark it before they come to the first workshop so as they are ready to apply the CAIAT software during the workshop using this test's results.	None	Technical support. See the workshop's design at (Appendix 6)
	Analysing and discussing the first application tests.	Observations about how the participating teachers use the CAIAT software and to what extent they know how to use its outcomes in terms of IAT considerations.	Observation sheet during workshops. (Appendices 7 and 8)
Second Application and Workshop (administered to pilot sample only)	Second application test: The teachers to construct and administer another specific test for a specific topic with concentration on HCD items, for the same purpose of the first test.	None	Technical support.
	Analysing and discussing the second application tests.	Observations about how the teachers use the CAIAT software and to what extent they know how to use its outcomes in terms of IAT considerations.	Observation sheet during workshops. (Appendices 7 and 8)
	To measure the participating teachers' opinions about their experience with the project's tools, concepts and skills.	Teacher opinions.	Reflective self-report
5.Complementary Procedures	Lesson observation Content Analysis for the teachers' tests And Interviewing the teachers	- Teachers' actual practice in terms of teaching on HCD level. - Teachers' overall perception about the project's main	check lists of: - Lesson observation - Content analysis - Guide for open

		components: HCD and CAIAT concepts	response interview (Appendix 9)
	Case Studies	Collecting a hands-on data about the way teachers use the CAIAT software and understand how to deal with its outcomes in terms of interpreting their indications. It measures the entire project's outcome of effectiveness as well as adoption level of participants.	Observation, content analysis and interview. (Appendix 10)

5.2.6.1 Introduction

The sample teachers are gathered in a meeting for presenting the HCD/CAIAT project to them. A questionnaire/pre-test is given at the end of this introduction to evaluate their background in computers, HCD, and IAT (Appendix 2). This is a pre-test, whose results are to be compared with post-test's results after training, so that the efficacy of training is measured as a first goal and the participants' background is uncovered as a second goal. Anonymity of the respondent is the most important characteristic of the questionnaire and tests as quantitative data collection instruments. However, since there is a comparison that will take place between the pre and post-applications then there was a need to set up their administration without violating anonymity. I have fulfilled this need by the following arrangement. At the end of the pre-test, each teacher was given an envelope and asked to put in her/his paper inside and seal that envelop by themselves and then to write down her/his name on the envelope. These were kept until the post-test finished, then each teacher had been given her/his envelop and asked to open it up and take the pre-test paper and then stable it along with the post-test paper without writing names on any of the two. Afterwards and to secure that the two papers do not miss each other, I had given a number for each of the set of two exams, for each individual, and wrote it down on both tests' papers. Other questionnaires and the self-report instrument that had been carried out at the end of the project are also filled in anonymously but have not been subject to similar procedures because there are no comparisons needed for their analyses.

5.2.6.2 Training

It is worthwhile to highlight that "training serves to help individual staff members gain new perceptions to foster change" (Fitch and Kopp, 1990: 27). The second stage was training participants in three issues: HCD concept, IAT, and the CAIAT software (Appendix 3 explains the training course). The training is given during school hours lasting three days for the pilot sample. As a response to the pilot application's results, this period for training has been extended to five days for the main sample.

At the end of the course, sections 2 and 3 that represent the pre-test part of Questionnaire/Pre-test instrument, which has been administered in the introduction stage (Appendix 2), have been re-administered as a post-test (see Appendix 4) to evaluate the efficacy of the training course per se. furthermore, comparing the two tests' results (pre and post-tests) helps in measuring the short-term outcomes of the training.

5.2.6.3 Contingent Application

This research acknowledges the notion of evolving change in the educational context from a contingent approach to change, in which gradation of application, bottom-up change, and adaptation to uncertainty in the environment should be undertaken (Rondinelli 1983a; 1983b in Rondinelli et al. 1990: 18) and (Rondinelli et al. 1990: 3). In this respect, this project's vision is that teachers adopt the technique of item analysis as an instrument for PD. This is to take place optionally and through making the CAIAT software and correspondent training available. It is anticipated then that a culture of item analysis is to spread and be a common practice. This research introduces HCD/CAIAT as a new concept, skill, and type of practice within this contingent vision. To gain the anticipated benefit of the CAIAT, the teachers were advised to apply concepts and skills of the project individually by writing unofficial tests, i.e., those that do not count for the final mark (quizzes for example) and then to analyse them by the CAIAT. These 'experimental' tests' focus is HCD where, by generating many HCD questions for a specific topic, more learning is expected to be gained. This stage is supposed to measure the level of interest that teachers could have for improving their ability in writing better HCD test questions and to measure the level of interest to analyse their tests by the CAIAT. At this point, they are not supervised or guided during this stage, therefore I called it a 'contingent' stage. During this stage, the

work team role is to support teachers who may face unexpected difficulties or need some help. At the end of this stage, a questionnaire was distributed to find out the level to which they were interested in taking advantage of the CAIAT to self-learn about their own abilities (Appendix 5).

5.2.6.4 Mandatory Application

5.2.6.4.1 First Application

Participants may and may not do the above sort of drill and practice routine; however, they were asked (mandatory) to construct and administer one specific test for a specific topic with concentration on HCD items. This test is an 'experimental quiz' as explained above. Thereafter, it became the focal content of the first workshop, which was held at a computer lab for all participants where there were two sessions. At the first session, they were asked to use the CAIAT software, which is available at the lab computers, to analyse their tests. At the second session, they were asked to form discussion groups, or focus groups, to learn from each other's experience and opinions and (Appendix 6). This sort of practice accommodates the notion of collaborative staff professional development much stressed by the related literature (Fitch and Kopp, 1990: 45). Moreover, it is to let the attending observers to record any ideas, comments, preferences, or difficulties they might reveal.

During their work on the CAIAT software, at the first session, the role of the researcher and/or his assistants is to observe the teachers for measuring how many mistakes they make, how fast they learn, and what main questions they ask (the form that is used by each observer is shown by Appendix 7). To make this task attainable, the teachers were divided into groups of ten so that each observer concentrates on ten individuals only. To facilitate this setting, the teachers were sitting in rows at the computer lab; each row is consisting of ten PCs. What is recorded on the observation sheet will be summarised latter on the summarising sheet (shown by Appendix 8). This entails looking at each observer's sheet row by row. Each row is describing one item of the observation task and is recorded by the observer for the ten teachers under observation in ten columns by two responses' scale (yes/no). The summarisation is carried out by counting which of the two responses (Y/N) is most for a given row across all observers' sheets; then the corresponding row at the summarising sheet will be ticked accordingly.

Unlike examinations, observation offers an optimum opportunity to: follow up behaviour of the participants in receiving the notions that will be presented; interact with trends that will be disseminated throughout the course; and record difficulties in understanding and obstacles in application, or ideas arisen during discussions. Observation also is intended to highlight the levels of ability that the teachers could show during practice sessions. Quality of this observation is achieved by the pre-designed form of authentic evaluation that examines each aspect of ability and leaves room for any other unlisted ones. The observation sheets as shown on Appendix 7 are intended to provide this level of quality.

5.2.6.4.2 Second Application.

This was administered for the pilot sample only. They were asked to draw on their first workshop findings and construct another 'experimental' test. Then, the second workshop with a similar setting of the first was held to see what they have gained from the former experience for the sake of this one. The research assistants are expected to benefit from repetition of application where, by this sense, observation during the second workshop is expected to be more focused on critical points and crucial aspects, which enhanced measurement validity.

This stage was not administered with the main sample because of two reasons. First, it is intended to inform the pilot study in the first place because it is considered the stage of the researcher's learning. Second, the main sample is of a greater number of individuals, which makes having them in another meeting for a second application more difficult. At the end of this stage, whether after the second application for the pilot sample or after the first with the main sample, each teacher was asked to write a reflective report demonstrating his/her experience.

5.2.6.5 Long-term measurements

5.2.6.5.1 Assessing the teachers' post-performance and further opinions

Appendix 9 illustrates the instruments that were utilised for this sort of assessment. However, the 'long-term' expression does not mean that it is a measurement that was applied on a long period enough to measure institutionalisation but a long-term measurement within the scope of time of this research fieldwork. therefore, it is a relative expression compared to the short-term measurements that are taken before, during or just after the training. Lesson observation comes on top of the

long-term procedures where the educational supervisors, who are the research assistants, arranged for a special visit to their teachers after the training and application stages were concluded. In these visits, and by a checklist shown on Appendix 9, they focused on the extent to which the teachers were utilising what they have trained for in terms of teaching on the level of HCD. They also looked at the teachers' tests and evaluated these tests in the light of the related checklist shown on Appendix 9. These visits are not different in nature from the regular visits the educational supervisors do during their routine work except that the teacher is asked to do some tasks pertaining to the research (the experimental HCD tests).

I needed to interview some of the participants to enrich my understanding of their interaction with the project and to uncover any critical points that needed to be understood. I did not address all of the issues that the interview would cover in advance because it was mainly intended to include any emerging issues during application and/or any points that were not presented sufficiently by the other instruments. However, there were initial issues that were set up as a groundwork that the interview started with as they appear on the guide shown in Appendix 9. I carried out some of these interviews with the pilot sample because they are in boys' schools. My assistants participated in the two samples as well. However, it is important to indicate that the purpose of this interview is to enrich the inquiry therefore, unlike other instruments; it will not have direct reflection to findings data. What it reveals will be used whenever it fit and/or feed the discussion.

5.2.6.5.2 Case Studies

Case studies are of great importance for collecting hands-on data about the way teachers use the CAIAT software and deal with its outcomes. It gives rich and reliable findings, and thus has been administered at the end of the project after all stages have been concluded. This is to let case studies measure the final outcome of the entire project. Data collection instruments that were used are observation, content analysis, and interview (Appendix 10). Observation was mainly used for discovering the extent to which the studied teacher is capable of using the CAIAT software, as well as for getting overall appraisal of the school's atmosphere. Content analysis is used for evaluating the test that the teacher has made for the CS in order to judge her opinion towards the test items upon what is being found by the expert evaluator. Interview is used at the end to

discuss with her about her test results in the light of item analysis findings that she has got by the CAIAT.

I mentioned earlier that I have selected a purposeful sample of 5 teachers for CS treatment. These five cases are female teachers therefore the leader of the supporting female team is the one who carried out the case studies. She has carried out many observations and interviews beforehand and hence have a first-hand experience in this term. However, she was enthusiast to learn more and develop her ability. She used the written style for recording responses during the CS treatment.

Scenario of the CS includes that the participant teacher provides the research assistant of a copy of her test including the questions sheet, pupils' answers and marks. After this, the teacher runs the CAIAT software, enters the marks and produces item analysis report without any help from the research assistant but records what she observes. After getting the analyses report, she has to comment on the strength or weakness of each item? What does she suggest for the weak ones? The research assistant is supposed to ask her further questions based upon her responses.

5.3 Research Dimensions

This research has two types of dimensions: the first type relates to inputs of the project (and could be also called themes or content objectives) and the second relates to outcomes. The first type consists of two dimensions: HCD and IAT. The HCD part has two areas: background knowledge of HCD concepts and skills on writing HCD questions. In practice, the first area is represented by the teachers' background basic knowledge about HCD and their abilities to reflect this in the behavioural instructional objectives that they write for their daily lesson plans; and the second is examined in two major areas of application: questioning during instruction and written questions in tests. The IAT part is represented by two areas: teachers' background knowledge of and skills in IAT and their ability to run/implement the CAIAT software effectively for this purpose.

I have to highlight that the last dimension is the prime dimension for the study; therefore, many research activities are centred on it. As mentioned before, problems in test questions are more likely to appear with those of the HCD levels, thus IAT practice's real need is in this arena. Actually, it should be noted that the HCD dimension of this project is working for the optimal fulfilment of the IAT dimension. Without the teachers knowing much about HCD concepts and questioning skills, training them about

IAT or providing them with a tool like the CAIAT to help them to utilise IAT easily may be a waste of time. This is why I regard this combination as one interventional package.

The second type that represents outcomes of the project consists of two dimensions: *effectiveness* and *adoption*. They represent the axes of the following chapters' presentation where the text is presented in sections that are based on these two dimensions. This classification comes from the fact that any project needs to be *effective* in order to be *adopted*. Input dimensions will be discussed within these outcome dimensions. *Effectiveness* is centred around the outcomes of the training of this project as represented by two types of outcomes: the immediate outcomes that appear just after the training (hence will be measured by the post-test and observation of the workshops just after the training course), and the long-term outcomes that are embodied in the teachers' skills during their classroom instructional practice, their written tests and their use of the CAIAT software after a period of time separating measurement from training. These will be measured by observation, content analysis, and case study. *Adoption* is represented by the level of interest that the teachers have in using the CAIAT software as well as the extent to which they pay attention to HCD questions during instruction or assessment practices. This is measured quantitatively by the teachers' self-reporting and qualitatively by observations and interviews during the fieldwork. The 'Effectiveness' is considered the prime dimension of this study since it is the most tangible aspect it deals with.

5.4 Research Questions

5.4.1 Main Questions

From the aim of this research, two main questions representing the mentioned above two outcome dimensions are formulated as follows:

1. To what extent can the HCD/CAIAT project assist female science teachers in Saudi schools improve their ability in analysing their test questions for writing exemplary HCD test items and teaching on HCD levels? (‘Effectiveness’ dimension).
2. To what extent could this be reflected in their on-going practice both for test construction and teaching? (‘Adoption’ dimension).

5.4.2 Sub Questions

Table 5.3 below lists the sub questions upon the research's two dimensions. For each dimension of the research, I am going to write a more specific sub questions that could be answered by data collection instruments' outcomes. Most of these will be two sections, one for examining the level of the targeted skill or observation and the other for examining significance of statistical differences according to the researched characteristics of teachers (the research variables). These characteristics are as mentioned in Table 5.4. However, the group of characteristics for adoption dimension has three additional characteristics more than those of effectiveness dimension namely: prior experience in using computers, possess a PC at home and ability to use some mainstream software packages. The reason is that these three relate to the extent to which the teacher has a good experience in computers; which is very likely to be seen as a catalyst to adoption. Table 5.4 summarises the questions that tackle the examination of statistical significance.

Table 5.3: List of the study sub questions

Dimension	Research Sub Questions
Effectiveness	**HCD content objective** Q1. To what extent do the researched teachers have a background of HCD concepts? Q2. To what extent can the researched teachers acquire skills in understanding HCD concepts? Q3. To what extent do the researched teachers have a background of writing HCD questions? Q4. To what extent can the researched teachers acquire skills in asking/writing HCD questions? **IAT content objective** Q5. To what extent do the researched teachers have a background in IAT? Q6. To what extent can the researched teachers acquire skills on IAT? Q7. What level of ability can the researched teachers acquire from the training of the project in general?
Adoption	**HCD content objective** Q8. To what extent will the researched teachers adopt what the project implies about HCD instructional objectives? Q9. To what extent will the researched teachers adopt what the project implies about asking/writing HCD questions? **IAT content objective** Q10. To what extent will the researched teachers adopt what the project implies about the CAIAT software and IAT skills?

5.4.2.1 Effectiveness Dimension

This dimension's questions seek to explore to what extent science teachers at Saudi schools can learn HCD and the CAIAT easily in order to apply what they have learnt in their work. For each question that includes examining statistical differences, a null hypothesis is derived for the purpose of examining its possibility by corresponding statistical treatment. Teachers' characteristics that such questions will be based upon represent the independent variables of each treatment; they are namely: educational qualification, prior training on test construction skills, prior training on IAT, key stage (intermediate/secondary), level of graduation, years of experience, and specialisation subject (physics – chemistry – biology).

Q1. To what extent do the researched teachers have a background of HCD concepts?

a. What level of ability can the teachers initially show about HCD concepts?

b. With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).

Hypothesis 1b

Statistically, there are no significant differences between the teachers' background experiences of HCD concepts that are associated with any of their researched characteristics.

Q2. To what extent can the researched teachers acquire skills in understanding HCD concepts?

a. What level of ability can the teachers acquire about HCD concepts as an immediate outcome of the project's short-term training?

b. With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).

c. What level can the teachers reach in acquiring skills pertaining to HCD concepts as an outcome of the entire project?

Hypothesis 2b

Statistically, there are no significant differences between the teachers' levels of acquisition of HCD concepts that are associated with any of their researched characteristic(s).

Q3. To what extent do the researched teachers have a background of writing HCD questions?

a. What level of ability do the teachers initially show in writing HCD questions?

b. With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).

Hypothesis 3b

Statistically, there are no significant differences between the teachers' background experience of writing HCD questions that are associated with any of their researched characteristics.

Q4. To what extent can the researched teachers acquire skills in asking/writing HCD questions?

a. What level of ability can the teachers acquire for asking/writing valid HCD questions as an immediate outcome of the project's short-term training?

b. With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).

c. What level can the teachers reach in acquiring skills of asking/writing valid HCD questions during instruction and within TA's testing as an outcome of the entire project?

Hypothesis 4b

Statistically, there are no significant differences between the teachers' acquisition for writing valid HCD questions as an immediate outcome of the project's short-term training that are associated with any of their researched characteristics.

Q5. To what extent do the researched teachers have a background in IAT?

a. What level of ability can the teachers show initially about IAT?

b. With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).

Hypothesis 5b

Statistically there are no significant differences between the teachers' background in IAT concepts and skills that are associated with any of their researched characteristics.

Q6. To what extent can the researched teachers acquire skills on IAT?

a. To what extent can the teachers run the CAIAT software for obtaining IAT parameters as an immediate outcome of the project's short-term training?

b. To what extent can the teachers apply IAT main concepts for evaluating their testing items as an immediate outcome of the project's short-term training?

c. In terms of IAT skills, with which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).

d. What level can the teachers reach in acquiring skills for using the CAIAT software as an outcome of the entire project?

e. What level can the teachers reach in acquiring IAT skills for evaluating their testing items as an outcome of the entire project?

Hypothesis 6c

Statistically, there are no significant differences between the teachers' acquisition of IAT skills that are associated with any of their researched characteristics.

Q7. What level of ability can the researched teachers acquire from the training of the project in general?

a. Is there any improvement when comparing the overall pre- and post-test results?

b. Can this improvement (if any) be interpreted by the training factor?

Hypothesis 7b

Statistically, there are no significant differences between the teachers' pre-test results and post-test results which are associated with the training of the project .

5.4.2.2 Adoption Dimension

This dimension is to explore to what extent the teachers adopt the CAIAT on their own initiative for improving their ability in writing better HCD questions.

Q8. To what extent will the researched teachers adopt what the project implies about HCD instructional objectives?

a. To what extent do the teachers think that the project has an impact of encouraging them to increase their use of HCD instructional objectives?

b. With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).

Hypothesis 8b

Statistically, there are no significant differences between the teachers' use of HCD instructional objectives that are associated with any of their researched characteristics.

Q9. To what extent will the researched teachers adopt what the project implies about asking/writing HCD questions?

a. To what extent do the teachers think that the project has an impact on encouraging them to increase their questions at the HCD level?

b. With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).

Hypothesis 9b

Statistically there are no significant differences between the teachers' practices of asking questions at HCD level that are associated with any of their researched characteristics.

Q10. To what extent will the researched teachers adopt what the project implies about the CAIAT software and IAT skills?

a. To what extent will the teachers use their own initiative in applying the CAIAT software?

b. To what extent will the teachers' professional development be elicited by the CAIAT software?

c. In terms of applying the CAIAT software on the teachers' own initiative, with which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).

Hypothesis 10c

Statistically, there are no significant differences between the teachers' use of the CAIAT software on their own initiative, by a self-learning, that are associated with any of their researched characteristics.

Table 5.4: Summary of questions that tackle examination of statistical significance

	Sub Question	The studied ability/skill
The teachers characteristics (study variables) for *effectiveness* dimension: educational qualification, prior training on test construction skills, prior training on IAT, key stage (intermediate/secondary), level of graduation (GPA or equivalent), years of experience in teaching and specialisation subject (physics – chemistry – biology)		
Effectiveness	***Sub Question Statement*** To what extent the researched teachers:	***Hypothesis Sub Question Statement*** With which variable(s) of the study differences between teachers (if any) are associated? ***Hypothesis Statement*** Statistically, there are no significant differences that are associated with teachers' researched characteristics and these differences are studies between the teachers':
	Q1. have a background of HCD concepts?	background experience of HCD concepts (Hypothesis 1b)
	Q2. can acquire skills on HCD concepts?	level of ability for acquisition of HCD concepts (Hypothesis 2b)
	Q3. have a background of writing HCD questions?	background of writing HCD questions (Hypothesis 3b)
	Q4. can acquire skills on HCD question construction?	acquisition of writing valid HCD questions as an immediate outcome of short-term training (Hypothesis 4b)
	Q5. have a background of IAT?	background of IAT concepts and skills (Hypothesis 5b)
	Q6. can they acquire skills on IAT?	acquisition of IAT skills (Hypothesis 6b)
	Q7. What level of ability they acquire from training in general? a. Is there improvement when comparing pre and post-tests overall results? b. Can this be interpreted by the training factor?	Statistically, there are no significant differences between teachers' pre-test results and post-test results that are associated with training. (Hypothesis 7b)
Teachers characteristics (study variables) for *adoption* dimension: educational qualification, prior training on test construction skills, prior training on IAT, key stage (intermediate/secondary), level of graduation (GPA or equivalent), years of experience in teaching, specialisation subject (physics – chemistry – biology), prior experience in using computers, possess of a PC at home and ability to use some mainstream software packages.		
Adoption	To what extent teachers can adopt what the project implies about:	***Hypothesis Statement*** Statistically, there are no significant differences that are associated with teachers' researched characteristics and these differences are studies between the teachers':
	Q8. HCD instructional objectives?	use of HCD behavioural objectives for instruction (Hypothesis 8b)
	Q9. writing HCD questions?	questions of HCD level (Hypothesis 9b)
	Q10. The CAIAT software and IAT skills?	initial use of the CAIAT software on a self-learn basis (Hypothesis 10c)

5.5 Data Collection Instruments

5.5.1 Overview

From the above illustration, it is worthwhile to summarise that the main data collection instruments are:

1. Questionnaire/Pre-test at the introductory stage.
2. Performance/Post-test after the training course.
3. Self-evaluation Questionnaire for the teachers after the contingent application stage.
4. Observation at all of the application stages.
5. Judges evaluation by comparing pre/ post-tests and the 1st and 2nd application stages' tests.
6. Reflective reports written by the participants, after the final application stage, on a form of an open-ended questionnaire.
7. Lesson observation for the participating teachers by their educational supervisor targeting their utilisation of what they have learnt about teaching on HCD level.
8. Content analysis of the teachers' tests.
9. Interviews with some of the participating teachers.
10. Case study

Distribution of the research's sub questions on these instruments is shown in Table 5.5.

Table 5.5: Research's sub questions and the corresponding data collection instruments

Data Collection Instrument	Sub Question
Questionnaire/Pre-test (Given to teachers at the beginning of introduction stage) Q1 is represented by Section1 of the instrument. Q3 is represented by Section 2 of the instrument. Q5 is represented by Section 3 of the instrument.	Q1a, Q1b, Q3a, Q3b, Q5a, Q5b, Q7a, Q7b.
Performance Post-test (Given after training course) Q2 is represented by Section1; Q4 and Q6 are represented by Section 2.	Q2a, Q2b, Q4a, Q4b, Q6a, Q6b, Q7a, Q7b.
Observation (during workshops)	Q6a
Lesson Observation (during instruction)	Q2c and Q4c
Content Analysis of the teachers' tests.	Q4c
Questionnaire to the teachers after the contingent application stage.	Qs: 8-10
Case studies	Q4c, Q6d, Q6e

The questionnaire is characterised by anonymity for the sake of collecting actual opinions. Observation and interviewing do not provide such a characteristic, however,

their importance embodied from the ‘in-depth’ investigation they offer. Consequently, the construction and design of both instruments would be dynamic to permit the addition of more illuminating questions just before or during the application. This intention follows the so-called 'sequencing approach' method combination (Fielding and Schreier, 2001). As a quantitative instrument, testing provides a sort of hands-on experience to reflect the level of the teachers' ability throughout specific stages of the project. The implementation of lesson observation and other qualitative data instruments fulfils the integrity of both traditions, as a cross-validation for potential findings. More theoretical elaboration on this will follow.

Besides the data collecting instruments, the project will use two important instruments: HCD/CAIAT training course package and the CAIAT software. Because the training course details are less relevant to the aims of this chapter, it is explained in Appendix 3. Also, the CAIAT software package concept has been explained on Chapter 4 while Appendix I provides elaboration on its features.

5.5.2 Quality of Instruments

5.5.2.1 Triangulation for validity

Although triangulation is a term that is used mainly in the discourse of the qualitative approach of research and is represented by the use of different data collection instruments, I am using it here with additional meaning of triangulating combined quantitative and qualitative method outcomes. In this regard, triangulation by the use of different methods, is a way to verify validity for the collected data. Fielding and Schreier presented a good illustrative comment on the different visions for triangulation:

> The term "triangulation" has acquired so many meanings and usages that it is now safer to use the terms "convergence" or "confirmation" when seeking cross-validation between methods (Fielding and Schreier, 2001).

Thurmond (2001), and some others, consider that triangulation is the combination of at least two methods. However, other opinions see it only more than one such as Cohen et al.'s (2007, p. 141) definition as an “attempt to map out, or explain more fully, the richness and complexity of human behaviour by studying it from more than one standpoint.” They categorised it into five types among these is the methodological triangulation which includes using more than a method for the same object of the study.

Campbell and Fisk (1959) consider this type as a check on validity (As cited in cohen et al. 2007). The present research will utilise this approach of triangulation.

McFee G. (1992) questioned triangulation's strength indicating that the value of its use in educational research is easy to "overestimate." He explained that the two types of methodological triangulation cannot achieve what they claim. He argued that the first method cannot guaranty that the various methods used tackle the same issue necessarily, and that the second method "fails to provide the sort of mutual support integral to the metaphor of triangulation." Also, Thurmond (2001) see that triangulation is used to decrease the deficiency of a single method. However, she concluded that triangulation does not strengthen a flawed study and stressed that, when used, it should be utilised for understanding the phenomenon. Nevertheless, Cohen et al. (2007, pp. 141-142) described the advantages of triangulation as "manifold," elaborating only two of them. The first includes that because human behaviour is complex, depending on one method is very likely to cause bias in researcher's investigations, and the second includes that each method has its bounds and limitations, therefore, using more than one method for measuring the same phenomenon overcome the problem of "method boundedness." In this respect, the triangulation between the quantitative and qualitative approaches represents an exemplary triangulation of this type. In addition, Cohen et al. (2007, p. 143) summarised that triangulation is suitable

> when a more holistic view of educational outcomes is sought (such as school effectiveness), when a complex phenomenon requires elucidation, when an established approach yields a limited and frequent distorted picture, or when a CS is conducted.

As seen above, this research design acknowledges the trend of multi-instrument usage, which provides a sort of cross-validation data collection method. The first main research question seeks to find out the functionality of the project, which makes it the main question of this study; therefore more sub-questions have been driven out to illustrate this questions' answer, along with six instruments (questionnaire/pre-test, post-test, observation, case study, and content analysis) that are employed in a triangulation basis for this sake.

The second question, however, is illuminating in order to explore the possibility of an effective outcome of the project in terms of the teachers' adoption. For this sort of question, I am adopting a contingent qualitative approach of inquiry, in which I am not seeking to find out direct effects of the project in this respect, but rather attempting to

explore the degree of probability for the anticipated outcome. Also, various methods/instruments of data collection are utilised to cross validate each other.

5.5.2.2 Reliability

Using SPSS© software, the reliability coefficient (Alpha indicator) will be calculated for instruments 1, 3, and 6. Responses of the other instruments are of a form that does not follow a similar statistical treatment for reliability; therefore, I will consider verification of validity for those instruments to be an adequate substitute in this regard where validity encompasses reliability (Brown, 1970: 98-99) and (Stanley and Hopkins, 1978: 114).

5.5.2.3 Validity

The following paragraphs will demonstrate validity for each of the mentioned instruments, covering the validity definition and verification procedures that I am going to follow.

5.5.2.3.1 Questionnaire

The questionnaire is the main tool for survey research. Therefore, it is worthwhile to have an overview of the questionnaire as a data collection tool, its characteristics, and validity with some elaboration on *self-reporting questionnaire*.

Questionnaire Overview

The major factor of the questionnaire that provides validity is its anonymity, while reliability is mainly achieved by the questionnaire's items being the same for all respondents. However, the main shortcomings of questionnaires are

> their lack of flexibility; low response rate, lack of control over environment, … no recording of spontaneous answers; difficulties in separating bad addresses from non-responses; no control over date of response; complexity of questionnaire format; and increased possibility of biased sample (CERIS, 2004: 26).

Nevertheless, the larger the sample, the more likelihood of questionnaire's reliability; hence the wide use of questionnaire in research since it provides a very useful instrument from big samples.

Questionnaire Validity

Because the questionnaire has a major role as a data collection instrument for this research, I am going to follow a cross-validation method for validity verification by examining this from various aspects of it.

External Validity refers to the notion of whether the study's results could be generalised for the rest of its population (Cohen et al., 2000: 109). In this respect I will adopt the positivist view, which insists that sampling should follow these three considerations: (a) a randomized mechanism of choice; (b) settings of application should avoid sources of bias; (c) the design should follow the standardisation procedural technique. For the first, this study targets the whole state's female science teachers and considers Al-Ahsa female science teachers as a 'typical' purposeful sample representing the group as a whole. The major goal of randomization is representation of teachers' characteristics related to their abilities and knowledge, thus the sample is fulfilling this condition of generalisation from qualitative research perspective. For the second, the way this questionnaire will be administered will acknowledge the anonymity requirement of respondents. For the third, standardisation will be fulfilled by piloting the questionnaire and finding out its reliability coefficient (Alpha), items' reliabilities, and internal consistency coefficient.

Content Validity is the extent to which the items on the questionnaire have covered all of the different aspects and issues regarding the main subject. It might be impossible, though, to do this by absolute scale, especially with the need to use a huge number of items that cover every possible entity of the studied phenomenon; but at most, there should be a satisfactory coverage of the main points that inform the aims and main questions of the study (Abo Hatab and Othman, 1976: 96). Questionnaires are designed on two forms: open and closed forms. The former consists of open questions that allow the respondents to write their opinions without any of the limits of closed questions. This is to aid in uncovering a variety of issues and interests that the researcher was not aware of. This could help her/him in the design of the closed form questionnaire, taking into account the findings of the open questionnaire, which helps to achieve content validity. In this study, this is fulfilled through the number of open questions that are included in the questionnaires, the reflective reports that come on a form of open-ended questionnaires and the interviews with some of the participating teachers. After these were administered to the pilot sample, their findings were analysed

to find out what issues were commonly reported by the respondents. These were added to the corresponding instrument for the main sample's application.

Face Validity is represented by a critical review that could be carried out by experts for the provisional version of the questionnaire. It is a kind of subjective validity that could be increased statistically by increasing the number of expert reviewers. If most of these experts commented that an item needed to be edited or deleted, then the researcher follows their recommendation if it fits the research purposes. Educational supervisors from both boys' and girls' educational authorities were considered experts for evaluating the questionnaire along with pre/ post-tests; where here were a number of comments, suggestions, and conservations as shown on Appendix 12.

Open-ended Questionnaire

Open-ended questionnaire is also called open-response questionnaire and is to give respondents the chance to express their own opinions, thoughts, feelings, and suggestions about what they have been/or will be part of. It is one of the best qualitative approaches for data collection. The open-ended questionnaire provides this facility with full anonymity, especially with small sample sizes. Further, as Cohen et al. (2000: 255) state:

> It is the open-ended responses that might contain the 'gems' of information that otherwise might not have been caught in the questionnaire…[It] can catch the authenticity, rich, depth of response, honesty and candour which … are the hallmarks of qualitative data.

Self-Reporting questionnaire

The main questionnaire questions in this study are self-reporting items, thus the following elaboration. According to Bill (1977), 26 previous studies found that self-reporting questionnaires were at least as valid as interviews. However, since a number of these studies lack external criteria and design quality because of comparing relatively unstructured interviews with highly structured questionnaires, Bill's research examined the relative efficiency of individual interview and self-completion questionnaire. He found a positive relationship that structured self-reporting questionnaires were "responded to at least as validly as were structured individual interviews" (Bill, 1977). As discussed by Huston and Robins (1982) and Metts, Spercher, and Cupach, (1991), self–report questionnaire has its limitations and problems such as: "respondents' limited awareness of their own thoughts, feelings, and behaviour; social desirability and self-serving bias." However, self-report questionnaire on the other hand, provides some

positive and important aspects such as measuring behaviour that take place during different times and places/situations. It also measures behaviour "retrospectively" (Noller, P. & Feeney, 2004).

Ferris's (1987) meta-analysis of a number of studies found that the self-reporting related to job characteristics has problems "less serious than initially believed" (Winefield, 2003). According to Harris & Schaubroek (1988) as well as to Conway & Huffcutt (1996), meta-analyses show that self-reporting (Or self-assessment) moderately relate to supervisors' and peers' assessment (Jones & Fletcher, 2004). Jones & Fletcher (2004) indicated that some studies have reported that the quality of self-reporting is improved when "instructions are framed in relative terms" (Farh & Dobbins, 1989; Mabe & West, 1982). They also indicated to Fox, Caspy, and Reisler's (1994) report that a non-balanced scale containing more positive than negative response options resulted in greater validity than a conventional scale of equal number of negative-positive items. By systematically evaluating a number of measurement conditions using self-assessment, Jones & Fletcher (2004) confirmed that "acceptable validities can be achieved [from self-reporting] but only when certain measurement conditions are employed. Specifically,…unbalanced questionnaires with motivational instructions."

5.5.2.3.2 Interview

Interviewing provides a framework within which respondents express, by their own terms, their understanding of the investigated subject. It also permits flexibility and diversity where different kinds of information such as opinions, feelings, knowledge, can be collected. Interview will take place in two research activities: case studies and the open response interview that will be applied as needed during the workshops and after lesson observation. *Focus groups* is a sort of interview and carried out through a guidance of moderator who provokes a talk between participants with one another and then that moderator records the revealed main points (CERIS, 2004: 24). Focus groups will be carried out in this study at the workshops during panel discussion of groups of acting participants.

Interview Validity

Validity of qualitative data collection instruments are difficult to be measured by procedural methods, such as those used for quantitative data collection instruments.

However, validity could be achieved by providing conditional characteristics and considerations before and/or during the application of these instruments (Best, 1981: 190). For an interview these considerations are:

1. Although an interview is characterised by its capacity for spontaneous and situational questions, basic questions should be prepared very carefully with a meaningful plan of introducing them in order (Best, 1981: 205).
2. There are different sources of bias; when they are eliminated, validity aspect is more likely to exist; among those are:
    - Recording bias; the way the interviewer records the interviewee responses may affect what he or she is recording: speed, concentration, and rephrasing, etc.
    - Clarity of questions; because words and sentences can take on different meanings. (Aqeel, 1999: 189).
    - The way of presenting questions; a specific introduction to the question may guide the respondent to a specific answer.
    - The interviewee extra comments should not steer the interview away because this may take his concentration from the main goal and may reflect on successive answers (Olayyan, 2001: 111; Aqeel, 1999: 188).
    - Avoidance of critical or difficult questions.
    - Encouraging the interviewee to respond freely by different ways: eye contact; friendly relationship; choice of time or venue etc. (Aqeel, 1999: 192).
3. Moral considerations such as respect for the interviewees, preserving their privacy, or cooperating with them if they have any kind of disabilities, and so on (Aqeel, 1999: 188-194).

It should be noted here that I did not carry out the interview because I cannot mix with female teachers, which might be interpreted as a threat to the validity primarily. However, my vision is different wherein I think that whether the researcher is to carry out the interview or somebody else is not a source of validity threat because what we are aware of is bias. Researcher's bias is more likely to offend the findings than his/her assistants' bias. The researcher's agenda, preferences, circumstances and gains/losses balance are more likely to be existed than the assistant's. This is because the researcher has tangible connection to the research that is based on gains and priorities of the organisation that he/she works for or the goal they are aiming to achieve.

5.5.2.3.3 Semi Structured Observation

Frese and Zapf, 1988, Glick et al., 1986 and Karasek & Theorell (1990, As cited in Winefield, 2003) outlined that observation could show distortion and bias. A number of problems underling Observation were highlighted:

1) limited time observation;
2) Unobservability of mental processes;
3) Effects of observation on work behaviour;
4) Halo and stereotyping effects;
5) Representativeness of workplace to be observed.

Semi structured observation will be undertaken during the workshops by the aid of a form (Appendix 7) including a list of skills or behaviours to be observed during the teachers' use of the CAIAT software (It is summarised by the form shown by Appendix 8). Also, it will be applied in lesson observation that will take place after the training and application stage by means of a checklist as shown in Appendix 9. *Structured Observation* is very systematic and enables the researcher to generate numerical data from observations. This indicates the quantitative nature of structured observation; however, a semi-structured observation may apply with qualitative approach considerations. As a qualitative instrument, validity for semi-structured observation could be fulfilled by similar considerations as mentioned above for interview. However, other considerations should be taken into account; among those are:

- Giving attention to major points to gain accuracy.
- Being natural so that the observed will not be offended.
- Recording, if needed, should not be noticed.
- Best not to participate, but if needed, then participation should be minimal.

These formed a kind of code of practice that was given to the observers before the workshop and the lesson observation. Fortunately, they are familiar with these considerations because, as educational supervisors, their job requires so.

5.5.2.3.4 Case Study

Many well-known case study (CS) researchers such as Robert K. Yin (1994), Robert E. Stake (1995), and Helen Simons (2009) have written about CS research showing its importance as a unique qualitative research method. It focuses on a limited number of events, therefore, CS can add strength to what is already known through previous research (Soy, 1997). Because CS is used to inform the two outcomes' dimensions of the present research showing findings about the entire project, I have elaborated in presenting the following theoretical background about CS.

Case study research refers to two research approaches: an in-depth study of a particular student, classroom, or school and the application of quantitative research

methods to non-probability samples (Postlethwaite, 2005). According to CERIS (2004, p. 61), a CS is "a detailed study of any kind of bounded system of interest - a project, a service, a participant." A key researcher in this area, Robert K. Yin (1994), defines the CS research method as

> an empirical inquiry that investigates a contemporary phenomenon within its real-life context; when the boundaries between phenomenon and context are not clearly evident; and in which multiple sources of evidence are used (Yin, 1994)

Technically, a CS is the optimal choice for answering the 'why' question. For example, a statistical survey might show that more pupils understand physics better in lab lessons, but it is case studies of a narrow group that will determine why this is so. It also answers questions of applicability. For example, a great computer application might be designed to encourage pupils to spend more time learning, but it is only by trying it out in a real life CS that will best reveal what evidence confirms the initial assumption. The present research's utilisation of the CS subscribes mainly to this notion.

Shuttleworth (2008) believes that for many years the CS was regarded as a valid method in social science and highlighted a number of characteristics of CS such as the realistic response it provides compared to that driven by purely statistical survey, and its flexibility in the sense of possibility to get unexpected results. He outlined that in CS the researcher is an observer rather than an experimental and that numerical data are utilised to judge trends not to draw solid conclusions.

As samples of convenience are usually employed (Postlethwaite, 2005), it is usually stressed that CS is not utilised for generalisation purposes (Yin, 1994; TESOL, 2003; CERIS, 2004; Postlethwaite, 2005; Shuttleworth, 2008). However, Yin (1994) commented that it could be used to "generalize to theoretical propositions, not to population as in statistical research." It is considered useful in providing qualitative information that clarifies previously obtained quantitative data (CERIS, 2004), which identifies the CS as a means of synergising collected data gathered by the two different approaches. Shuttleworth (2008) indicated the importance of synergising the two approaches through case studies whereby their findings are "tied in with more general statistical processes." In the present study, this trend is adopted by utilising CS findings in order to derive evidence-based data that confirms what is found by other quantitative statistical methods.

Stake (1994) have outlined three types of CS. the first is *intrinsic* CS in which one case is the target to be studied and the research is concerning that case per se. The second is *instrumental* CS, in which the interest is not the studied case but a phenomenon that this case's behaviour/interaction will reveal. The third is *collective* CS, which is similar to the second type but using multiple cases for data collection. The present research utilises CS of the third type.

Data gathered by CS is normally qualitative, but it may also be quantitative. Various instruments could be utilised for collecting data such as interviews, observation, surveys, or document analysis (Soy, 1997). Badr (1989) considers that the interview is the most important instrument for CS data collection. The prime interest here is triangulation of data collection so that the various methods of data collection consolidate each other which is suggested to increase construct validity (Yin, 1994). This is because CS is subject to bias by the researcher's implementation or interpretations. The researcher should carefully observe and identify factors associated with the observed phenomenon. Also, in interviews, addition of questions may be necessary as the study progresses (Soy, 1997). Training on how to apply CS has a vital role in harvesting validity. In this respect, five investigative skills are highlighted: question asking, listening, adaptiveness and flexibility, grasp of the issue being studied, and lack of bias (Bickman & Rog, 1997). In the present study, observation, content analysis, and interview are the instruments used for the study's CS. Validity of CS comes from validity of these instruments which has been achieved through piloting and reconstructing. Training was not necessary because the research assistant who have undertaken the case studies is well trained in interviewing and observation as I have mentioned earlier. However, some readings were given to her in advance so that she becomes more informed about CS requirements such as flexibility, dynamic question asking, investigative observation to unexpected findings, so on.

There is a necessity to follow a formal CS “protocol” as an important tactic for upholding quality control for CS research. The protocol defines rules and procedures that will be followed in data collection and is especially stressed when the CS is a multiple-case design so that consistency is assured. (Bickman & Rog, 1997). In the findings chapter, the sort of protocol followed in executing the CS method across the five cases is explained showing that consistency was an obligational aspect for this process. Also, Chapter 10 lists the main questions that steered the CS's different data collection methods used: content analysis, observation and interview.

5.6 Statistical Analyses

The questionnaire will be standardised by finding out reliability coefficient (Alpha), items' reliabilities (that are found by calculating the overall reliabilities after item deletion), and internal consistency coefficient, which represents statistically the inter-correlation amongst the questionnaire's items or test scores. Some findings will be presented by means of statistical significance indicators after administering a T-Test and ANOVA test for the corresponding items. This is to show significant statistical differences (if any) among different groups of teachers; these are classified upon: educational qualification, prior training on test construction skills, prior training on IAT, key stage (intermediate/secondary), level of graduation, years of experience, specialisation subject (physics – chemistry – biology), prior experience in using computers, possession of a PC at home and ability to use some mainstream software packages.

Qualitative data, on the other hand, will be analysed and then demonstrated in either a percentage form whenever this applies; or by narrative presentations in which justifications, inferences, reasoning, logical relationships, and evaluations will be presented according to what the collected data reveal, and upon congruency of their inclusions with each other on one hand, and with the quantitative findings on the other.

I am going to use the following statistical analyses for the collected quantitative data:

a. Descriptive Statistics

These include percentage ratios, frequencies, means and graphs.

b. T-Test for Independent Samples (Two-Tailed Module)

This is used to show whether there are any significant statistical differences in the studied point that are associated with a specific variable of a dichotomous value as it states in sub questions 1-5b, 6c, 8c, 9b, 10b.

c. T-Test for Paired Samples (Two-Tailed Module)

This is used to show the significance of statistical differences between pre/post-tests in order to discover the extent to which training has fulfilled its aims. This is carried out for question 7b.

d. Analysis of Variance (One-way ANOVA) Test

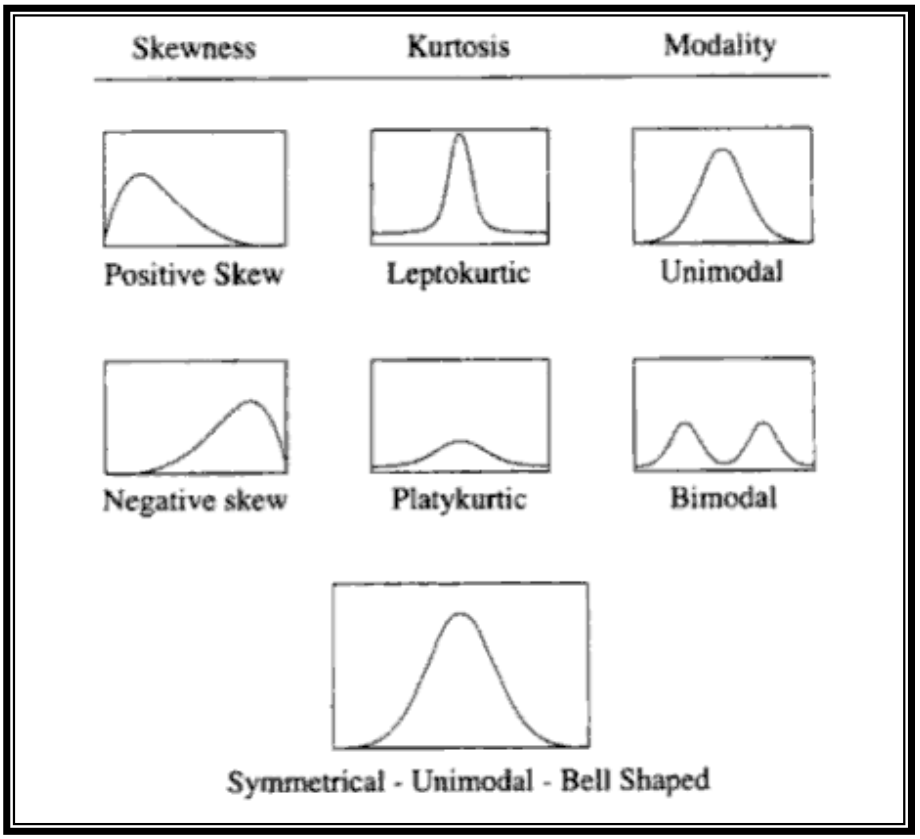
This is used to show whether there are any significant statistical differences in the studied point that are associated with a specific variable that has grouped data of more than two values as it states in sub questions 1-5b, 6c, 8c, 9b, 10b.

5.6.1 Statistical Analysis Methods

Percentages and histograms are the two main methods of descriptive statistics that are used for presenting the quantitative aspects of findings. A percentage is calculated by finding the ratio of the number of respondents for a specific selection/value of the item with respect to the total number of the sample. However, the total is not fixed for all cases since there are missing responses sometimes, thus the number of those that respond is the reference in such cases.

The histogram that has been used is the bar chart accompanied by a curve sometimes for the purpose of referencing them to the corresponding standard curves, such as symmetrical normal distribution (Bell shaped) or bimodal curves. Newton and Rudestam (1999: online) have illustrated six shapes of normal distribution curves as they appear in Diagram 5.1 below. The bimodal curve is distinguished for representing achievement tests results with two peaks. The first peak represents the lower achieving students and the second peak represents the higher achieving students.

Diagram 5.1: Possible shapes of normal distribution (Newton & Rudestam, 1999)



Raw data of the pilot sample are shown in Appendix 11. These were used to find quality indicators for the data collection instruments used for the research before being utilised with the main sample of the study. For examining reliability of the questionnaire I have used Guttmann split half and Alpha Cronbach reliability coefficients. For examining its internal consistency as a statistical validity indicator I have used item inter-correlation matrix. By examining the significance of statistical differences between or among cases that the T-Test and ANOVA methods were utilised, the statistical analysis with Levene and Sheffe tests supported them respectively. For applying these analyses to my data, I have used SPSS© for Windows© and the Release 11.0.0 standard version. In the following paragraphs I will further explain these indicators and will highlight the obtained values for quality of the data collection instruments. Appendices 13-15 show SPSS© outputs of these indicators' statistical calculations.

5.6.1.1 Correlation

Correlation is defined by Ross (2005) that it "involves the search for relationships between variables through the use of various measures of statistical association." This relationship does not necessarily entail causality, but can give an indication of its likelihood. The researcher cannot depend conclusively on correlation coefficient to address causality, but needs to find other indicators or evidence that prove its existence, if at all. Some other statistical techniques that can be utilised in this respect are factorial analysis and variance analyses. As an indicator of the relationship between studied variables that helps researchers to gain better understanding of the studied phenomenon, correlation is used very frequently in research. Moreover, its concept is utilised within other statistical parameters to provide more comprehensive or sophisticated indicators (Factor Analysis, Point Bi-serial, Split half, etc.). In the present research, correlation is not part of the parameters used for interpreting or discussing findings but is utilised in calculations of: (a) reliability: by Split-half and Alpha coefficients; (b) validity (inter consistency): by inter correlation matrix and the non-parametric correlation Spearman co-efficient.

5.6.1.2 T-Test Method

This test is used for comparing two means in order to show whether the difference between them is as significant from a statistical point of view, or in other

words to see whether they are ‘real’ differences or just have come by chance (Ghareeb, 1985: 316). It is the preferred test for examining small samples that include less than 30 cases (ibid, 1985: 321). However, modern T tables can fit to big samples that reach up to 10,000 cases (Assyed, 1978: 335). There are four types of this test and this study has used two of them.

a) Independent Samples Two-Tailed T-test

This is used for comparing means of two samples that are independent from each other and have been subjects for the same treatment. They could represent a different number of individuals, i.e. n1#n2. Two-Tailed stands for the case in which the researcher does not assume a prior anticipation of any difference, and the aim is to find if there is any difference of a significant value (Assyed, 1978: 338). In this study, this test is used for finding significant differences between the means of the two groups that have been divided upon one characteristic of the group's individuals, which should be of a dichotomous nature such as their educational qualifications.

b) Paired Samples Two-Tailed T-test

This test is similar to the previous one, but the two samples are to be of the same characteristics, or they could be one sample that actually has been subject to two different treatments such as pre-/ post-tests. In this study, this is the case where pre and post-tests results were compared by means of this type of analysis.

5.6.1.2.1 Conditions of using T-Test

There are four conditions that should be met in order for T-test use to be valid. The size of the sample should be more than 5 cases, differences between the two samples should not be very high, distribution of the two samples should be a normal one, and there should be homogeneity of the two samples as measured by comparing the two variances by Levene's test. Since all of these conditions are being examined by SPSS© software, which I used for calculating T values, therefore, no need for examining these. However, Levene's test appears in the initial tables of SPSS© output from which I select only the needed information for the findings' presentation. For Levene's results, SPSS© prints out a table for T-test analysis as in the example shown below in Table 5.6, then I will choose one of the resulting T-test values upon the value of significance of

Levene's test. If the first row “Equal variance not assumed” value was significant then the second row value would be chosen and vice versa (Najjar, 2003: 69).

Table 5.6: Independent samples test

After attending the training course, have your questions in the level of HCD, either during instruction, in worksheets, in home works or in your tests, increased?	**Levene's Test for Equality of Variances**			**t-test for Equality of Means**						
		F	**Sig.**	**T**	**Df**	**Sig. (2-tailed)**	**Mean Difference**	**Std. Error Difference**	**95% Confidence Interval of the Difference**	
									Lower	**Upper**
	Equal variances assumed	2.155	.152	-.767	32	.449	-.1172	.15291	-.42867	.19424
	Equal variances not assumed			-.735	22.177	.470	-.1172	.15957	-.44798	.21355

5.6.1.3 Analysis of Variance (One-way ANOVA) Test

This test is used for comparing means of more than two samples that are independent from each other and have been subjects for the same treatment. They could be of a different number of individuals, i.e. n1#n2#n3, etc., and have been divided upon one characteristic of individuals such as teachers' levels of graduation, their years of experience, and subjects of specialisation. Sheffe test is used then to show which group of the studied variable's value the differences can be correlated to.

5.6.1.4 Split-half Coefficients for Reliability

Split-half method is the indicator for reliability of a questionnaire in which items of the questionnaire are split into two parts and the correlation coefficient between the two is calculated as an indicator of these items reliability, or the whole scale reliability. Among the different models of calculating this coefficient, I have selected Guttmann's model because it assumes that variances of the two halves are not equal (Abdulrahman, 1983: 205) which is suitable for this study's questionnaire. According to the output report of SPSS© which is presented in the appendix 13, the resulting reliability coefficient value by Guttmann's split-half model for the pilot sample is 80%, which is a good value, among the 35 cases involving male teachers.

5.6.1.5 Alpha Coefficients for Reliability and Internal Consistency

Alpha coefficient (Cronbach) is an indicator that is calculated for a questionnaire's reliability and for showing the degree of its items' internal consistency by measuring average inter-item correlation (SPSS, 2001). This could also be considered as an indicator of the construct validity. I have applied this calculation to items 1, 2, 3, 4, and 5 of the questionnaire. These are the main questions of the questionnaire that represent its major function and all teachers should respond to each of these items, while the partial questions within these items, vis. q4a, q4b, q4c, etc., are to be answered conditionally and thus were not considered as part of the scale's components. According to the output report of SPSS© which is presented in Appendix 14, the resulting Alpha reliability coefficient for the pilot sample (35 cases) is 0.6869 (69 %). This is a good value as a reliability coefficient for the scale; but, looking at the internal consistency dimension through the list of "Alpha if Item Deleted" shows that item number 1 has caused a decline in the whole scale reliability since its corresponding value in this list is 0.7288, which is greater than the scale's Alpha; which also means that the scale's Alpha would be of greater value if item 1 had been excluded. Moreover, correlation matrix at the same appendix shows that this item's correlation column is the only one amongst the scale's items that has weak correlations, which indicates that it might not be relevant to this group of items.

This is the only item that has shown less relevance to the scale which: (a) indicates the value of the rest of items being fit to the scale; (b) indicates the need to re-calculate the scale's Alpha after excluding item 1, which I have done. Appendix 15 shows these results reporting that the corrected Alpha reliability coefficient, after deleting item 1 from calculations, is 0.7353 (74%), which is a good value for a scale of 5 items only. Looking at the correlation matrix in this analysis shows that correlations between the items have improved as well; this indicates a good statistical validity level for these items. Nevertheless, this does not necessarily call for cancelling item 1 completely from the questionnaire since it has an important function for the information it seeks to collect. The calculations of reliability after deleting item 1 are useful for indicating quality of the rest of items as well as the whole instrument. However, in this sense, item 1 could be considered similar in nature to the fundamental items of the questionnaire, vis. number of years of experience, educational qualification, etc.

Because the statistical test: non-parametric correlation Spearman coefficient using 2 tail method shows the level of significance of correlation between items, it was applied to these five items and Appendix 16 shows the correlation matrix. Item 1 is the lowest in its correlations to other items, which follows the same discussion above. Item 2 correlates best with item 5 of a significant value of 57%. However, correlations with items 3 and 4 (26% and 25% respectively) although not significant, I think are within a satisfactory level of relationship when looking at their content, where item 2 is about using the CAIAT software while items 3 and 4 are about using HCD instructional behavioural objectives and HCD questions respectively. Item 3 have significant correlations to items 4 and 5, which is logical as they look at related issues. Item 4 correlates significantly to item 5

5.6.2 Missing Data Treatment

According to Howell (2007) there are many treatments to missing data. Among these: *Listwise* deletion, in which, cases with missing data are omitted. It leads to unbiased parameter estimates but has the disadvantage of reducing the sample size. Missing data in my work are in pre/ post-tests results and they have two forms.

The first form is that some marks were missing at one of the three parts of pre/ post-tests. My way to treat these is to consider them zero since they are test results for which not answering one part of the test means that it is very likely the tested individual does not have the answer. The second form of missing data is that some cases were of missing marks at all parts of the test (pre/post-test), which indicates an absence of those teachers at that test time. They were in low percentage not exceeding 10%, therefore, I chose the *Listwise* method of missing data treatment by omitting these cases and did the analysis without them.

5.7 Important Considerations

5.7.1 Limits of this research

This research is limited to Saudi schools in the city of Al-Ahsa, and covers only female science teachers at intermediate and high schools. It tackles only one aspect of the test construction problem as a contribution to other possible treatments, such as intensive cascade training, supervision, and management; thus, it is limited in this respect. It also suggests that improving teachers' ability to write HCD questions could

be achieved through an analysis of test items. This assumption does not claim that this research acts as a panacea for testing problems; rather, it is one of a number of important factors which might play a significant role in this area because “item analysis is no substitute for meticulous care in planning, constructing, criticizing, and editing items” (Stanley and Hopkins 1978, p. 268). Usually, IAT is used for standardised tests, not teacher-made tests, but this classical trend is behind this project's vision.

5.7.2 Limitations of this research

This study did not track the impact of its inclusion on pupils’ learning, despite its importance; because an impact on learning mostly appears after a long period of time (see the CASE project, for example). Therefore, finding out about such a relationship needs a longitudinal study, which is beyond the present work’s time scale. This study also relied on the research assistants for data collection, due to a number of reasons. First, some data collection procedures required that educational supervisors reflected on the extent to which the teachers were applying the training they had received. This could only be carried out by each teacher's educational supervisor so that their behaviour would be revealed in a natural situation. Second, the researcher is a male and the main sample individuals are female teachers working in girls’ schools, where males are not allowed to mix, according to the single sex policy of education provision in the KSA. Third, during the workshops and the training, the observations could not be carried out by only one person, since the number of participants was high. Therefore, the research assistants’ role in this respect was an inescapable necessity.

The adoption dimension depended, mainly, on the questionnaire that was given to the teachers at the end of the project to report their practice over the whole project period. It is characterised by being self-reporting, which some may criticise in terms of validity despite the fact that the self-reporting questionnaire is valued by many of the researchers discussed earlier. Nevertheless, when considering the other party’s opinion, I believe that assessing adoption by an in-depth measure that provides hands-on experience needs a comprehensive longitudinal research approach, which is beyond this work’s time limits and size. My estimation is that in such studies, the data collection could only be undertaken, at the very least, over two or three academic years. However, as triangulating qualitative data strengthens self-reported data, what this study has revealed in terms of adoption are indicators that will be very likely to encourage decision-makers to positively “try out” this opportunity with a high level of confidence.

5.7.3 Comment

I will indicate a number of observations in executing this effort during the fieldwork showing the lessons I learnt from these experiences. For example, when I introduced the work plan to the educational supervisors who will work with me as research assistants during the pilot phase (with male teachers), they raised some concerns. Initially, I was designing the training to be held after school and hence the teachers would be given allowances for this. Nevertheless, the educational supervisors commented that their past experience showed teachers had little interest in training at such a time, even when an allowance was paid, and that it made no real difference from their perspective. I appreciated their opinion as sources of expertise and changed the course time to the morning, during school hours. As for the main sample, the female teachers, certainly this concern applied, due to their circumstances as mothers and also having no flexible personal transport to take them to the training venue after school time[29].

The educational supervisors raised another obstacle to the training, which is the support of school head-teachers who mostly do not welcome such activities and prefer the teachers not to attend. This mostly comes from head-teachers' background knowledge about 'ineffective' training, especially if they do not know on the details of what the training will consist of. One solution for this was to invite the school head-teachers to attend a presentation of the project. We did that and fortunately, there was a surprising acceptance of and admiration for the project. Without the concern revealed by the educational supervisors I would not have paid sufficient attention to the head-teachers' role. This highlights for me the great importance of the researcher communicating with all the stakeholders and understanding the context in which she/he is going to work before jeopardising her/his plans and/or agendas. With the main sample (female teachers), a similar approach was followed.

On the first day of training the pilot sample (male teachers) we began by giving the teachers the first instrument, a questionnaire and a pre-test. This was planned to be done in 30 minutes, but took 90 minutes, which affected the following events in that day. This is because I decided to follow a new idea that ensures anonymity but this affected the time plan. I assumed that the idea would be understood easily by all, which was a false assumption. The lesson learnt here is that the researcher should think thoroughly about scenarios of fieldwork and never underestimate the time needed for

new ideas and unusual procedures, however simple they might appear initially to him/her. Doing case studies is another challenge, given the fact that Saudi people in the education sector are not used to case study research treatment. Moreover, training the designated research assistant for this task on doing the case study required me to read thoroughly about it and refer to my previous experiences especially the one that I carried out in an English school during my MA study in the UK..

I discovered that fieldwork was not as easy as I had imagined. Although I hold a high stake position in the educational system (being GM for GDGEA) I was not able to apply every piece of the fieldwork exactly as I had planned. People's circumstances, preferences, attitudes or understanding of the activity in process played a major role in framing the way the fieldwork could function. Furthermore, I can say that at the present moment I have a new vision of how to deal with fieldwork if I need to do so again. More pre-procedures will be applied as well as more time will be allowed for training on and crafting the use of the CAIAT software. Also, I would make use of pioneering teachers in training and following-up instead of educational supervisors whose role for improving teachers' performance could be intersecting the mission of this project.

Chapter 6
F i n d i n g s

6.1 Organisation of the Chapter

Connecting these to the research questions, I will present the quantitative findings in charts, curves and tables of data or through statistical indicators. These will be followed by some qualitative findings that are implemented for triangulating the quantitative data. Throughout this chapter and Chapter 7, and wherever appropriate, some additional qualitative findings such as the interview results, observation notes, work team members' comments, and any other not-planned-for results will be employed within my comments and explanations. To maintain a good flow of thoughts throughout the text, I have kept most of the technical explanations and those relating to the pilot sample in the 'Endnotes' section. I will present my main findings based on the two dimensions of the study: *effectiveness* and *adoption*. Throughout the tables, *F* stands for frequency and *P* for percentage. My comments on these data are for explaining what the numbers and percentages mean, leaving additional elaborated discussions to Chapter 7.

This research has been applied to two samples. The first is a male teachers' pilot sample with a sample size of 43 individuals and the second is a female teachers' main sample with a sample size of 409 individuals. However, due to a variety of reasons, some of the participants did not attend all sessions of the project, thus the total number of participants in the findings tables do not match each sample size.

To help in giving an overall vision of this chapter structure I provide Table 6.1, which shows the distribution of research questions over the study dimensions and the corresponding data collection instruments classified by quantitative and qualitative types of data. For sections that include hypotheses testing, different treatments are applied to examine the level of significance of the factors that could have an impact on each dimension. These types of findings are presented by Table 6.8 for factors related to the effectiveness dimension and by Table 6.19 for factors related to the adoption dimension. In each table, two groups of findings are presented: T-test and F-test findings, in which any presented number is a final value that represents the level of significance for the corresponding variable, leaving the other related numbers that produced these to be presented in detail by the Tables in Appendix 17. The minimum statistically significant level of confidence that I have chosen for examining my data is

95%, which means that the obtained significance values must be equal to or less than a P-value of 0.05 in order to be considered acceptable significant indicators.

Table 6.1: Research questions and their corresponding sources of data

Study Dimensions and Research Questions	**Data Collection Instrument**	
	Quantitative Type	**Qualitative Type**
Individuals' Basic Data	Questionnaire before the training course	___
Effectiveness Dimension		
Research Questions of HCD		
HCD Concepts		
Q1a	Pre-test results	___
Q2a	Post-test results	___
Q1b, Q2b	Statistical analysis	___
HCD Question Construction Skills		
Q3a	Pre-test results	___
Q4a	Post-test results	___
Q3b and Q4b	Statistical analysis	___
Q2c, Q4c	___	1) Lesson observation 2) Content analysis 3) Case studies
Research Questions of IAT		
Q5a	Pre-test results	___
Q6a and Q6b	Post-test results	___
Q5b and Q6c	Statistical analysis	___
Q6d and Q6e	___	Case studies
Research Questions about Functionality of Training		
Q7a	Comparison of Pre- and Post-test results	___
Q7b	Statistical analysis	___
Adoption Dimension		
Research Questions of HCD		
HCD Objectives		
Q8a	Questionnaire of contingent application stage	___
Q8b	Statistical analysis	___
HCD Questioning		
Q9a	Questionnaire of contingent application stage	___
Q9b	Statistical analysis	___
Research Questions of IAT		
Q10a	Questionnaire of contingent application stage	
Q10b	Questionnaire of contingent application stage	___
Q10c	Statistical analysis	___

6.2 Individuals: Basic Data Findings

I start with a demonstration of the main sample's participants' basic demographic information about issues that relate to the research dimensions. Table 6.2 shows this information. The table shows that participants come from the two key stage areas, intermediate and secondary, with nearly identical percentages that are distributed over three main areas of science specialisation with apparently a lesser number of candidates from the field of physics. Those shown as specialised in nutrition are actually biology teachers. The idea of asking the questions in item 4 of Table 6.3 and item 6 of Table 6.4 arose after the application of the pilot sample's fieldwork for enriching the profile of the participants' background. It should be noted that the findings of Tables 6.4, 6.5 and 6.6 are responses from individuals who answered 'Yes' to item 5 of Table 6.3. A further point to be noted is that these findings should not be compared to the findings of Table 6.3 directly. The purpose of Diagram 6.1 is to illustrate that the statistical distribution of the participants' levels of graduation is the normal distribution type, with its median shifted to a point above the middle (that is, more than 2.5).

Table 6.2: The main sample: teachers' basic data

#	Item	Response							
1	Graduated from College of Education?	**Yes**				**No**			
		Frequency		**Percentage**		**Frequency**		**Percentage**	
		246		75%		82		25%	
2	GPA	**Excellent**		**Very Good**		**Good**		**Satisfactory**	
		F	P	F	P	F	P	F	P
		60	17.8%	196	58.2%	73	21.7 %	8	2.4%
3	Years of Experience	**<= 5 years**		**6 – 10 years**		**> 10 years**			
		F	P	F	P	F		P	
		47	13.9 %	138	40.7 %	154		45.4 %	
4	Which key stage?	**Intermediate**				**Secondary**			
		F		P		F		P	
		179		51.7%		167		48.3%	
5	Which specialisation?	**Biology**		**Chemistry**		**Physics**		**Nutrition**	
		F	P	F	P	F	P	F	P
		135	38.9%	114	32.9%	90	25.0%	8	2.3%

Diagram 6.1: Distribution of levels of graduation

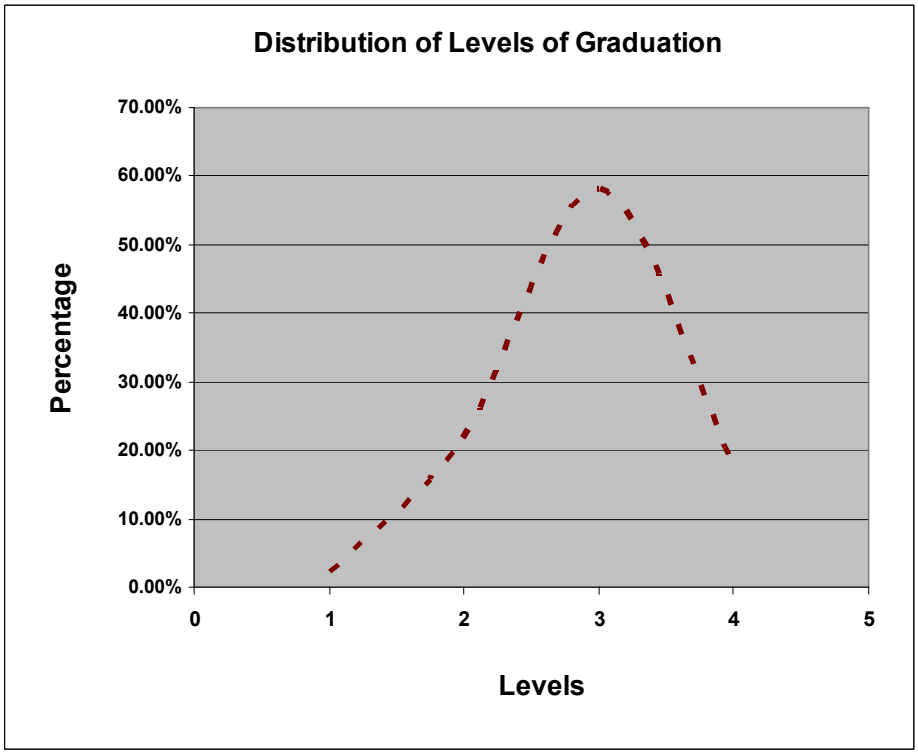


Table 6.3: The main sample teachers' general background on the projects' concepts and skills

#	Item	Response			
		Yes		No	
		F	*P*	*F*	*P*
1	When writing a test, do you make a specification table?	97	31.8%	208	68.2%
2	Have you ever attended a training course on test construction?	194	57.2%	145	42.8%
3	Have you ever attended a training course on test results analysis?	35	10.5%	299	89.5%
4	Considering the test results, have you ever noticed a problem in any of your test questions?	104	32.6%	215	67.4%
5	Do you know how to use a computer?	265	77%	79	23%

Table 6.4: Sub group percentage of installation of a software package

#	Item	Response			
		Yes		No	
		F	*P*	*F*	*P*
6	Have you ever installed any software package on a PC?	68	24.7%	207	75.3%

Table 6.5 below reveals that Microsoft Word© is the software most commonly used by the teachers (74.3%), which is congruent with common sense. Furthermore, this result is congruent with the percentage of item 5 in Table 6.3, which reveals that 77% of the participants know how to use a computer. In fact, further analysis of the results indicated that all of those who are computer users know how to use Microsoft Word©, which explains the high convergence of the two percentages and at the same time provides cross-reliability for these figures.

PowerPoint comes in second place, which also looks logical, since many teachers tend to use this software to present their lessons. Other software packages are less common, especially SPSS©. It should be pointed out here that the answer regarding Microsoft Explorer (8.1%) contradicts the answer of item 5 in Table 6.6. Findings there reveal that 34% and 15% are using a computer for browsing the internet either all the time or most of the time, respectively; in other words, the sum is almost 50% using internet much of the time, whilst the findings of Table 6.5 reveal that only 8.1% are using Microsoft Explorer. Bearing in mind that Microsoft Explorer is the most popular internet browser, this raises a contradiction. Nevertheless, the problem in my opinion is with the 8.1% figure, where many of those who did not choose Internet Explorer in this selection actually have no idea of the name of the browser they use. Interestingly, I have examined this potential by asking a number of students: what is the name of the ‘thing’ that you use for searching and browsing the internet and the majority did not give the right answer; some said “It is messenger” I think that it was not appropriate to use the term 'Microsoft Explorer' for this item but I had no reasonable alternative because it is the most popular software for this purpose and this question lists the names of software packages, not descriptions as the next question does.

Table 6.6 identifies that the common purposes for a candidate to use a computer are for browsing the internet, searching databases for non-personal purposes and word processing. The results indicate that using computers for decision-making tasks such as managing personal financial affairs or doing calculations is not common. These tasks are extremely relevant to the present project's task, which indicates the importance of training individuals on how to utilise computers in order to make decisions rather than how to use them per se.

Table 6.5: Areas of computer experience possessed by the main sample's participants

Name of Software	Percentage	Name of Software	Percentage
Microsoft Word®	74.3%	Microsoft PowerPoint®	41%
Microsoft Explorer®	8.1%	Microsoft Access®	7.8%
Microsoft Excel®	32.4%	Microsoft Outlook®	6.6%
SPSS®	0.3%	____	___

Table 6.6: The main sample participants' purposes for using a computer

No	Item	All the time	Mostly	From time to time	Rarely	I don't use a computer
		P (%)	*P (%)*	*P (%)*	*P (%)*	*P (%)*
1	As a word processor.	**32.2%**	**8.4%**	**9.5%**	**7%**	**31.9%**
2	To manage my personal financial affairs.	**3.6%**	**2.2%**	**10.6%**	**12%**	**71.5%**
3	For presenting (lessons, training courses, etc.)	**22.9%**	**12.2%**	**28%**	**9.7%**	**27.2%**
4	For scheduling tasks, appointments, etc.	**6%**	**3%**	**7.5%**	**13.5%**	**69.9%**
5	Browsing and searching the Internet.	**34%**	**15%**	**21.4%**	**10.5%**	**19%**
6	Doing calculations (other than personal).	**4.1%**	**3.4%**	**3.7%**	**15.3%**	**37.5%**
7	Searching databases for personal purposes (such as telephone directories, names or addresses of people, organisations or products, etc.)	**7.9%**	**5.1%**	**13%**	**13%**	**16%**
8	Searching databases for non-personal purposes (such as electronic encyclopaedias, dictionaries and reference books)	**16.7%**	**9.2%**	**18.1%**	**12.8%**	**34.3%**

6.3 The Findings of the Research Questions

6.3.1 Effectiveness Dimension

This dimension reveals two major issues: the level of readiness that the participants possess prior to the project training and the level to which they can learn and apply its underpinning skills. Presentation of the findings will be organised upon the project's two content objectives: HCD and IAT. However, the HCD part includes two areas: the background knowledge of HCD concepts and the skills in writing HCD questions. IAT is also represented by two areas: IAT knowledge/skills and the ability to

run/implement the CAIAT software for this purpose. Descriptive statistics such as average and median are collectively shown in Table 6.7. In addition, hypotheses testing findings are shown on Table 6.8.

Table 6.7: Descriptive Statistics for the Pre- and Post-tests

Statistic	HCD				IAT	
	Knowledge of HCD Concepts		Skills of HCD Questions		Knowledge of IAT	
	Pre-test	**Post-test**	**Pre-test**	**Post-test**	**Pre-test**	**Post-test**
Mean	4.447	7.921	4.583	6.375	4.988	8.127
Median	5	8	4	6	6	8
Standard Deviation	2.95	1.24	1.972	1.805	2.735	1.177
Kurtosis	-1.2	-0.05	-0.41	0.781	-0.61	7.692
Skewness	-0.23	-0.56	-0.11	-1.07	-0.8	-1.711
Range	10	6	8	8	10	10
Minimum	0	4	0	0	0	0
Maximum	10	10	8	8	10	10
Count	347	315	333	316	347	315

Table 6.8: Independent samples test (T values) and One-way ANOVA test (F values), Summary of *Effectiveness* Dimension [30]

	Targeting differences between individuals upon their:	**HCD**		**IAT**
		Knowledge of HCD Concepts	Skills of HCD Questions	Knowledge of IAT
		T Values Significance		
Pre-Test	Educational qualification	**0.036 ***	**0.844**	**0.235**
	Training on test construction	**0.037 ***	**0.942**	**0.008 ***
	Training on IAT	**0.013 ***	**0.343**	**0.075**
	Key stage (intermediate/secondary)	**0.037 ***	**0.934**	**0.648**
Post-Test	Educational qualification	**0.452**	**0.502**	**0.228**
	Training on test construction	**0.143**	**0.089**	**0.762**
	Training on IAT	**0.412**	**0.388**	**0.733**
	Key stage (intermediate/secondary)	**0.380**	**0.001 ***	**0.001 ***
		F Values' Significance		
Pre-Test	Level of graduation	**0.071**	**0.054**	**0.194**
	Number of years of experience in teaching	**0.984**	**0.384**	**0.825**
	Specialization subject (physics – chemistry – biology).	**0.032 ***	**0.000 ***	**0.954**
Post-Test	Level of graduation	**0.318**	**0.340**	**0.060**
	Number of years of experience in teaching	**0.054**	**0.193**	**0.196**
	Specialization subject (physics – chemistry – biology).	**0.210**	**0.083**	**0.229**

6.3.1.1 Effectiveness in HCD

6.3.1.1.1 Research Question 1

Q1. To what extent do the researched teachers have a background of HCD concepts?

a. *What level of ability can the teachers initially show about HCD concepts?*

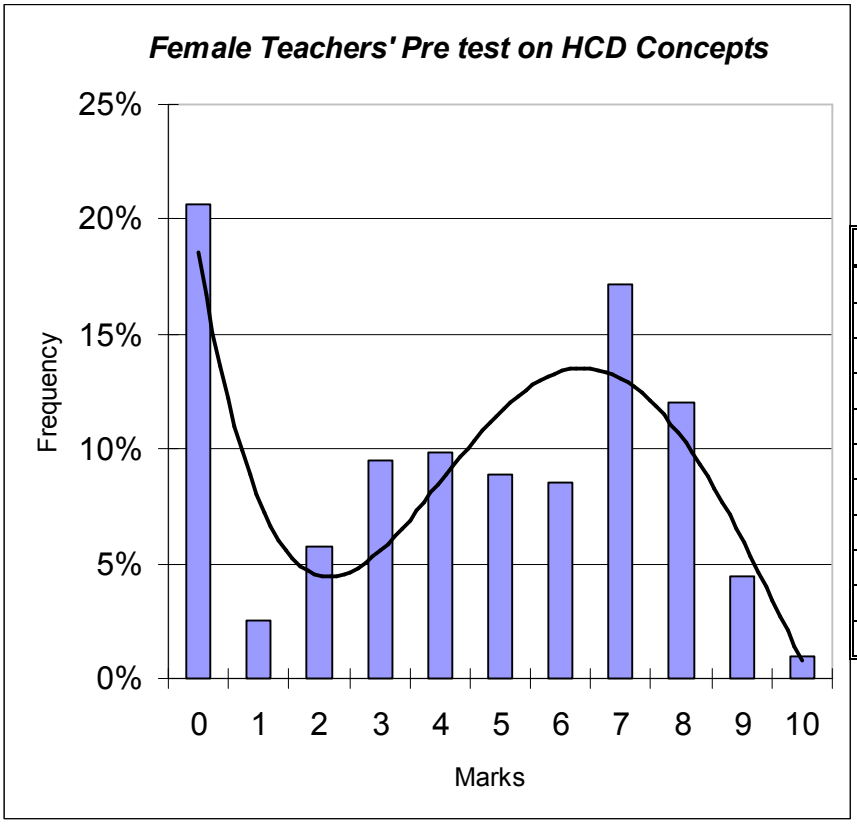
b. *With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).*

Hypothesis 1b

Statistically, there are no significant differences between the teachers' background experiences of HCD concepts that are associated with any of their researched characteristics.

Diagram 6.2 shows that teachers' background in HCD concepts is within a normal distribution, a form of bimodal curve, in which two peaks represent two groups of examinees: low and high achievers. Also, Table 6.7 shows that the mean is around 4.45 and the median is 5. On a test scale of 10, this indicates almost a mid-point, which confirms the already mentioned normal distribution. Question 1 is answered from this finding that the teachers did not show a previous high level of knowledge about the HCD concept because their test results showed a normal distribution. Therefore, this level is very likely to be described as being limited by the academic achievement that they acquired from pre-service training without having additional knowledge from a PD exercise. In fact, teachers are expected to perform much better when tested on these concepts that are fundamental to the teaching profession. This is stressed further when we take into account the basic level of this project's pre-test questions. I will continue with this notion in my appraisal for the findings to come that are of a similar nature.

Diagram 6.2: Distribution of participants' pre-test results in HCD concepts



Marks	Frequency
0	21%
1	3%
2	6%
3	10%
4	10%
5	9%
6	9%
7	17%
8	12%
9	4%
10	1%

Question 1b requires an examination of the significance of statistical differences between the teachers' results on the HCD concepts section of the pre-test. From Table 6.8, it can be seen that the T value for educational qualification is significant at 0.036 since it is below 0.05[31]. The same applies to the other three characteristics for the pre-test section of Table 6.8, which leads to rejecting the null hypothesis 1b for this table's pre-test characteristics. Thus there are significant statistical differences between teachers' background experience of HCD concepts according to (a) their educational qualification: these differences favour those with higher means; that is, those educationally qualified; (b) their previous training in two skills: test construction and IAT, which favour those trained; (c) their key stage, which favours those who teach in a secondary school.

The F values section of Table 6.8 shows that the first two characteristics have no significant F value at the 0.05 level, thus the null hypothesis 1b is accepted for these, which means that there is no association between teachers' background experience in HCD concepts and their level of graduation or number of years of experience in teaching. However, the subject of specialisation (physics – chemistry – biology) appears to have a significant F value, which leads to rejecting the null hypothesis 1b for specialisation and considering that there is an association which is explained by the Scheffe test results for this variable. This is shown in Table 6.9, which reports that test differences are significant between Biology and Nutrition in favour of Biology and between Physics and Nutrition in favour of Physics. This indicates that Biology and Physics are the two subjects most responsible for the differences between the teachers' results in their HCD concepts background.

To summarise, the answer to Q1 is that the teachers generally have a limited background of HCD concepts; however, it is at a higher level for those educationally qualified, those who have previous training on test construction or IAT skills, those who teach in a secondary school or those with a specialisation in Biology or Physics.

Table 6.9: Scheffe test results for specialisation variable

What is your specialisation (I)	What is your specialisation (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Biology	Chemistry	0.1885	0.37212	0.968	-0.8570	1.2340
	Physics	-0.1185	0.39812	0.993	-1.2370	1.0000
	Nutrition	3.0481	1.06455	0.044 *	0.0574	6.0389
Chemistry	Biology	-0.1885	0.37212	0.968	-1.2340	0.8570
	Physics	-0.3070	0.41252	0.907	-1.4660	0.8519
	Nutrition	2.8596	1.07002	0.069	-0.1465	5.8658
Physics	Biology	0.1185	0.39812	0.993	-1.0000	1.2370
	Chemistry	0.3070	0.41252	0.907	-0.8519	1.4660
	Nutrition	3.1667	1.07933	0.037 *	0.1344	6.1990
Nutrition	Biology	-3.0481	1.06455	0.044 *	-6.0389	-0.0574
	Chemistry	-2.8596	1.07002	0.069	-5.8658	0.1465
	Physics	-3.1667	1.07933	0.037 *	-6.1990	-0.1344

* The mean difference is significant at the .05 level.

6.3.1.1.2 Research Question 2

Q2. To what extent can the researched teachers acquire skills in understanding HCD concepts?

a. *What level of ability can the teachers acquire about HCD concepts as an immediate outcome of the project's short-term training?*

b. *With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).*

c. *What level can the teachers reach in acquiring skills pertaining to HCD concepts as an outcome of the entire project?*

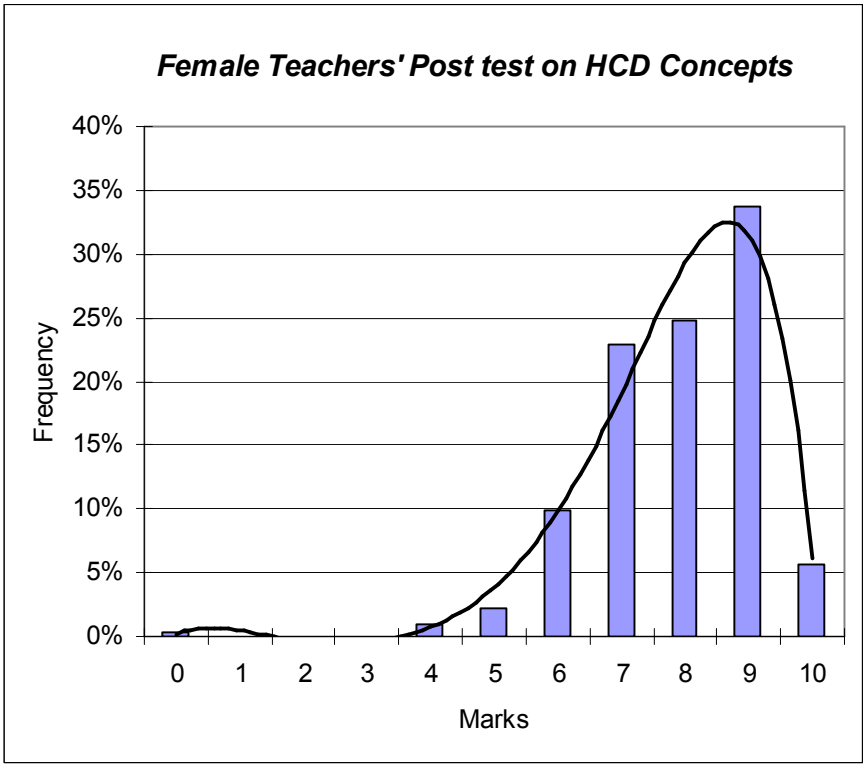
Hypothesis 2b

Statistically, there are no significant differences between the teachers' levels of acquisition of HCD concepts that are associated with any of their researched characteristic(s).

Diagram 6.3 shows teachers' HCD skills results after completing their training; the distribution is a negative skewed curve. Moreover, the level of change in the curve's skewness is reflected by the median pre- and post-values, which according to Table 6.7

have shifted from 5 to 8. Furthermore, the mean has risen from 4.5 to almost 8, which reflects the good level of increase in the trainees' abilities. The other statistics in Table 6.7 reflect similar results. Comparison of this curve to that of pre-test results in Diagram 6.2 reveals a good return from training, which answers what question Q2a asks about: that the training has resulted in increasing the teachers' abilities in relation to HCD concepts to a high level.

Diagram 6.3: Distribution of participants' post-test results on HCD concepts



Marks	Frequency
0	0%
1	0%
2	0%
3	0%
4	1%
5	2%
6	10%
7	23%
8	25%
9	34%
10	6%

Research question 2b requires the examination of the significance of the statistical differences between the teachers' results on the HCD concepts section of the post-test. These differences are categorised by the different variables shown on Table 6.8. A glance at the tables' results of the post-test sections show that all T and F values are statistically not significant at the 0.05 level, which leads to accepting the null hypothesis 2b, yielding an overall inference that the outcome of training the teachers in HCD concepts is independent from any factor of the study's variables.

To answer Q2c, lesson observations[32] that have been carried out in the teachers' lessons provides information about the extent to which the teachers have gained skills in writing valid instructional objectives for their lessons that are at the HCD level. During their school visits, the educational supervisors reported this observation by looking at

the teachers' lesson plans, and this represents an outcome of the whole project – not just the training course. There were 267 teachers who participated in the observation, representing 66% of the sample. Table 6.11 shows these findings in items 1 and 2 whose high percentages indicate that the teachers have acquired this skill to an excellent level. This is congruent with the project's immediate training outcome as mentioned above from the post-test's findings.

In summary, the answer to Q2 is that the project was effective in providing the teachers with knowledge and skills in HCD concepts both for short and long-term outcomes and that for the latter, it is independent from all teachers' researched characteristics.

6.3.1.1.3 Research Question 3

Q3. To what extent do the researched teachers have a background of writing HCD questions?

a. *What level of ability do the teachers initially show in writing HCD questions?*

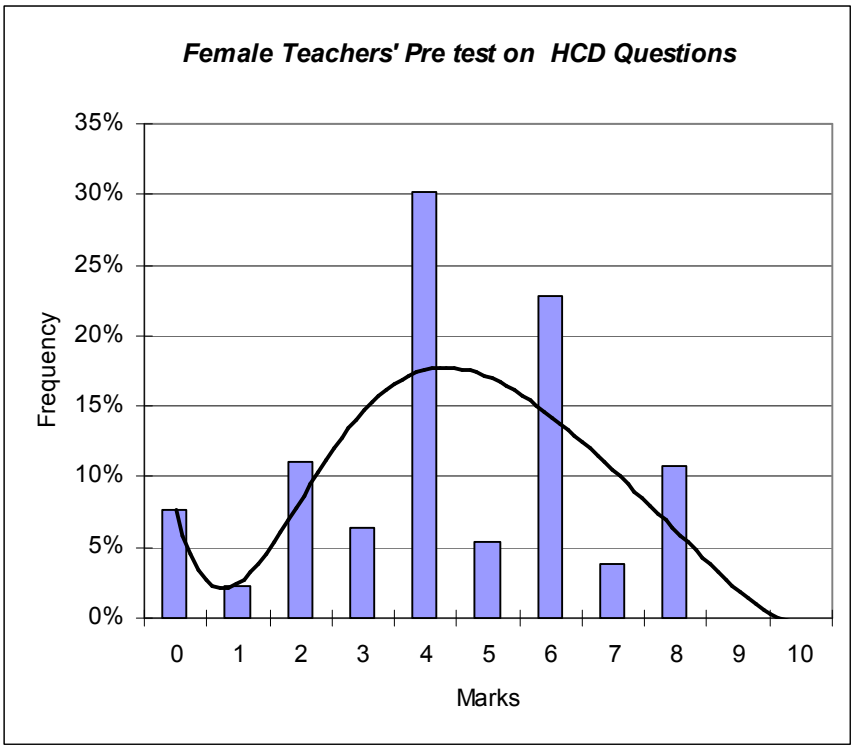
b. *With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).*

Hypothesis 3b

Statistically, there are no significant differences between the teachers' background experience of writing HCD questions that are associated with any of their researched characteristics.

Diagram 6.4 shows a slightly positive skewed curve, which indicates a limited previous ability in asking/writing HCD questions. In addition, Table 6.7 shows that the mean and median values are respectively 4.6 and 4. All of these indicate that the teachers have a limited competency of background skills in writing HCD questions; which answers Q3a.

Diagram 6.4: Distribution of participants' pre-test results in skills of constructing HCD questions



Marks	Frequency
0	8%
1	2%
2	11%
3	6%
4	30%
5	5%
6	23%
7	4%
8	11%
9	0%
10	0%

For Q3b, Table 6.8 shows that all teachers' characteristics of the pre-test section have no significant T and F value at the 0.05 level except the subject specialisation variable, thus the null hypothesis 3b is accepted for all variables and rejected for the specialisation variable, which means that the differences between the teachers' results on the pre-test for background experience in writing HCD questions could be interpreted based upon their subject specialisation only. Table 6.10 illustrates the related Scheffe test results in which these differences are reported between Biology and Chemistry in favour of Chemistry, between Biology and Physics in favour of Biology and between Chemistry and Physics in favour of Chemistry. This leads to an overall result that Chemistry and Biology specialisations are responsible for differences in teachers' background for writing HCD questions

In summary, the answer to Q3 is that the teachers have a limited competency for writing HCD questions. However, the level of competency is related to the teachers' specialisation in Chemistry or Biology and is not associated with any other researched characteristics of the teachers.

Table 6.10: Scheffe test for teachers' background in writing HCD questions

What is your specialisation (I)	What is your specialisation (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Biology	Chemistry	**-1.2776**	**0.23228**	**0.000 ***	**-1.9303**	**-0.6249**
	Physics	**0.8828**	**0.25010**	**0.007 ***	**0.1800**	**1.5856**
	Nutrition	**-0.3731**	**0.65312**	**0.955**	**-2.2084**	**1.4622**
Chemistry	Biology	**1.2776**	**0.23228**	**0.000 ***	**0.6249**	**1.9303**
	Physics	**2.1604**	**0.25893**	**0.000 ***	**1.4328**	**2.8880**
	Nutrition	**0.9045**	**0.65656**	**0.594**	**-0.9404**	**2.7495**
Physics	Biology	**-0.8828**	**0.25010**	**0.007 ***	**-1.5856**	**-0.1800**
	Chemistry	**-2.1604**	**0.25893**	**0.000 ***	**-2.8880**	**-1.4328**
	Nutrition	**-1.2559**	**0.66307**	**0.311**	**-3.1191**	**0.6074**
Nutrition	Biology	**0.3731**	**0.65312**	**0.955**	**-1.4622**	**2.2084**
	Chemistry	**-0.9045**	**0.65656**	**0.594**	**-2.7495**	**0.9404**
	Physics	**1.2559**	**0.66307**	**0.311**	**-0.6074**	**3.1191**

* The mean difference is significant at the .05 level.

6.3.1.1.4 Research Question 4

Q4. To what extent can the researched teachers acquire skills in asking/writing HCD questions?

a. *What level of ability can the teachers acquire for asking/writing valid HCD questions as an immediate outcome of the project's short-term training?*

b. *With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).*

c. *What level can the teachers reach in acquiring skills of asking/writing valid HCD questions during instruction and within TA's testing as an outcome of the entire project?*

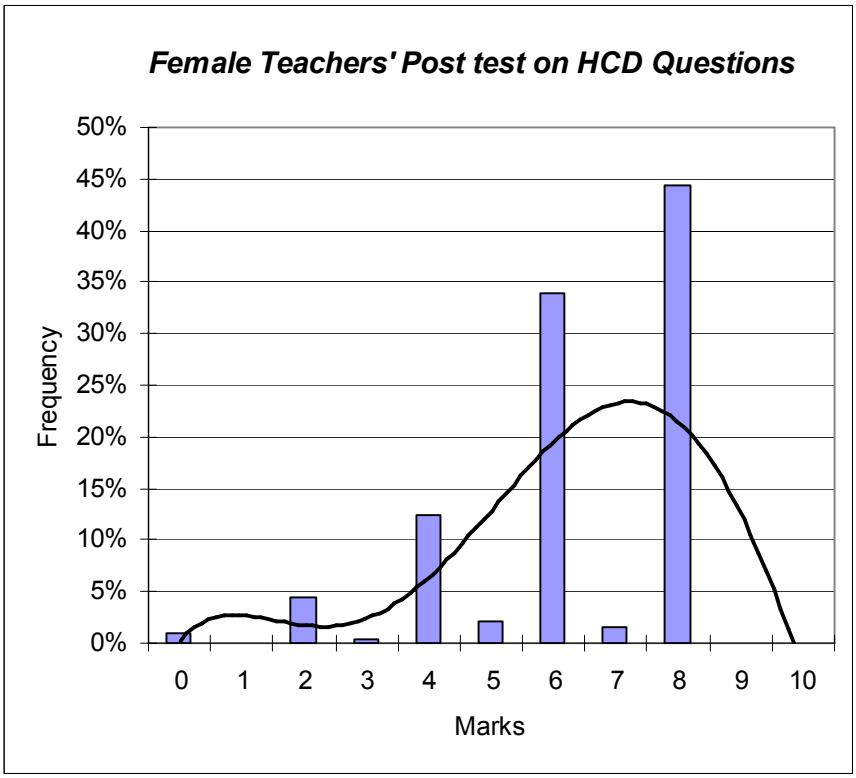
Hypothesis 4b

Statistically, there are no significant differences between the teachers' acquisition for writing valid HCD questions as an immediate outcome of the project's short-term training that are associated with any of their researched characteristics.

Diagram 6.5 shows the teachers' HCD questions background results after completing their training, and the distribution is a negative skewed curve which, when compared to that of the pre-test at Diagram 6.4, clearly reveals a fair return from

training. Moreover, Table 6.7 reveals that the mean value has increased from 4.6 to 6.4 and that the median value has increased from 4 to 6, thus confirming that improvement from the training for this skill has occurred. This answers research question 4a.

Diagram 6.5: Distribution of participants' post-test results in skills of constructing HCD questions



Marks	Frequency
0	1%
1	0%
2	4%
3	0%
4	12%
5	2%
6	34%
7	2%
8	44%
9	0%
10	0%

Table 6.8 shows that none of the teachers' characteristics of the post-test section have significant T and F value at 0.05 level except the key stage variable, thus the null hypothesis 4b is accepted for all variables and rejected for the key stage variable. This means that differences between the teachers' results on the post-test for acquiring skills on writing HCD questions can be only interpreted by the key stage factor in favour of those teaching in secondary school.

To answer Q4c, lesson observations that have been carried out in teachers' lessons and the content analysis of their tests provide further information about the extent to which the teachers have gained overall skills on creating HCD questions as an outcome of the whole project – not just the training course. The teachers who participated in the content analysis were 256, representing 63% of the sample. The findings from Table 6.12 show that from items 1 and 2 of Table 6.11, 92.5% of the sample teachers are asking HCD questions, with 97.2% of those asking good quality

HCD questions, and 85% of these good quality questions are thought to be challenging for the learners. Items 6, 7 and 8 show that the majority of the sample teachers are asking HCD questions and following good practice in this respect, whether in asking questions, receiving answers or reinforcing the pupils' responses. As an anticipated result of following HCD questioning, item 9 asks whether a teacher's related pedagogical practices are suitable for teaching thinking requirements. The result is that 89.9% of the sample teachers were fulfilling this requirement. Furthermore, 93.5% of the sample teachers' pupils were responding actively to this level of questions during instruction, and for 96% of the cases, the pupils were giving answers. These high results reveal that the teachers' practices in HCD questioning during instruction are high. However, this is triangulated by other measurements, as the following paragraph will show.

The teachers were asked to write a special test of HCD questions only. This was to examine the extent to which they could produce as many valid HCD questions as possible for a specific subject, which provides a hands-on measurement and shows what level of ability they reached. Table 6.12 shows that 91.4% of the tests included HCD questions. Out of these, 39.3% included no less than 80% of the test questions of HCD level, 17.3% included no less than 60% of HCD level. Summing up, these results reveal that 56.6% of the teachers who acquired the skill succeeded in writing no less than 60% of their tests with questions of HCD level[33]. This indicates a fair level of the skill attainment. However, it is not highly congruent with the higher percentages of attainment found by the lesson observations. I think that what the content analysis revealed is more likely to be representative of the actual level since it is a measurement that takes place on a consistent basis over a long period of time with the freedom to deliberate and scrutinise the information. Furthermore, some educational supervisors who took the two measurements might have carried out the lesson observations with less attention to accuracy, whilst their judgement of the content analysis was more critical because they considered that written materials could be looked at later by somebody else, hence this made them pay greater attention to content analysis. It is important to utilise a third type of findings which is that of case studies (CSs), which I will discuss shortly below. Table 6.14 summarises CSs findings where three out of the five studied cases show a good level in HCD question construction skills. This cross-validates the content analysis finding and hence confirms its result that the project was fairly effective in the HCD question construction dimension.

In summary, the answer to Q4 is that the project was effective in providing the teachers with skills in writing HCD questions both for short and long-term outcomes and that for the latter, it is independent from all teachers' researched characteristics except for the key stage where more acquisition has taken place by secondary school teachers.

Table 6.11: Results of teachers' lesson observations (HCD dimension)

#	Items	Yes		No	
		F	*P*	*F*	*P*
1	Do they use behavioural objectives from HCD levels?	244	91.4%	23	8.9%
2	Are these behavioural objectives stated well?	237	97.1%	7	2.9%
3	Do they ask HCD questions during instruction?	247	92.5%	20	7.5%
4	Are their HCD questions of good quality?	240	97.2%	7	2.8%
5	Are these questions eliciting some challenges amongst pupils?	210	85%	37	15%
6	Do they follow good techniques in asking such questions?	233	94.3%	14	5.7%
7	Do they follow good techniques in receiving pupils' responses?	230	93.1%	17	6.9%
8	Do they follow good technique in reinforcing pupils' responses?	233	94.3%	14	5.7%
9	Are their pedagogical practices suitable for teaching thinking requirements?	222	89.9%	25	10.1%
10	Are the pupils interacting actively with questions of this type?	231	93.5%	16	6.5%
11	Are pupils able to give answers?	237	96%	10	40%

Table 6.12: Results of content analysis of teachers' tests (HCD dimension)

#	Items	No of Cases	No of Missing Cases	Response					
				Yes			No		
				F	*P*		*F*	*P*	
1	Are the test questions at HCD level?	256	0	234	91.4%		22	8.6%	
2	What is the ratio of those that relate to HCD?	214	42	80%-100%		60%-79%		Less than 59%	
				F	*P*	*F*	*P*	*F*	*P*
				84	**39.3%**	**37**	**17.3%**	**93**	**43.5%**
3	Are there any defects in stating the questions?	218	38	45	20.6%	173	79.4%		
4	If there are defects (45 cases) are they of the type that IAT could indicate?	45	0	30	66.7%	15	33.3%		
5	Are there answers that indicate quality of learning?	208	48	148	71.2%	60	28.8%		
6	What is the percentage of this type of answer out of the total?	161	95	More than 30%		10%-30%		Less than 10%	
				F	*P*	*F*	*P*	*F*	*P*
				81	**50.3%**	**67**	**41.6%**	**13**	**8.1%**

6.3.1.2 Effectiveness in CAIAT/IAT

6.3.1.2.1 Research Question Q5

Q5. To what extent do the researched teachers have a background in IAT?

a. *What level of ability can the teachers show initially about IAT?*

b. *With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).*

Hypothesis 5b

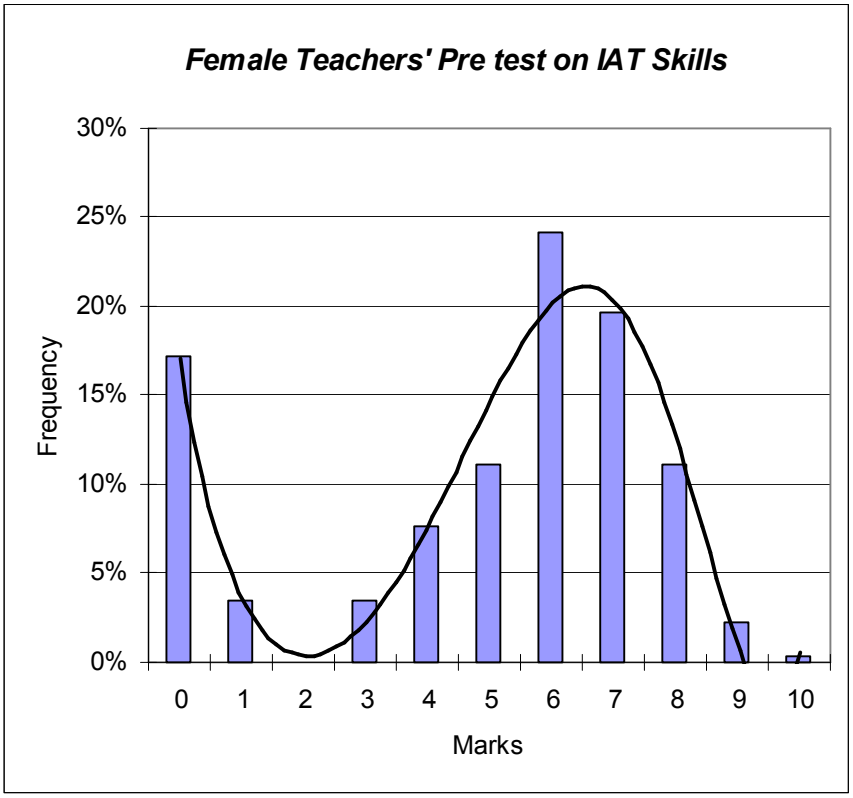
Statistically there are no significant differences between the teachers' background in IAT concepts and skills that are associated with any of their researched characteristics.

Diagram 6.6 shows that a teachers' previous background in IAT is within a kind of normal distribution which is similar to that of the HCD concepts as described above. Therefore, this answers Q5a that the teachers' previous knowledge of IAT is limited by the academic achievement that they acquired from their pre-service training.

For Q5b, Table 6.8 shows that all teachers' characteristics of the pre-test section have no significant T and F value at 0.05 level except for the training on test construction variable which means that the null hypothesis 5b is to be accepted for all variables and rejected for the variable of training on test construction which means that differences between teachers' results in the IAT section of the pre-test can only be interpreted by their prior training on test construction in favour of those trained.

The conclusion and answer for Q5 in the light of these findings is that the teachers have limited previous knowledge of IAT concepts and skills. However, more ability is shown by the teachers who have prior training in test construction, whilst the other teachers' researched characteristics have no association with this knowledge.

Diagram 6.6: Distribution of participants' pre-test results in IAT skills



Marks	Frequency
0	17%
1	3%
2	0%
3	3%
4	8%
5	11%
6	24%
7	20%
8	11%
9	2%
10	0%

6.3.1.2.2 Research Question Q6

Q6. To what extent can the researched teachers acquire skills on IAT?

a. *To what extent can the teachers run the CAIAT software for obtaining IAT parameters as an immediate outcome of the project's short-term training?*

b. *To what extent can the teachers apply IAT main concepts for evaluating their testing items as an immediate outcome of the project's short-term training?*

c. *In terms of IAT skills, with which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).*

d. *What level can the teachers reach in acquiring skills for using the CAIAT software as an outcome of the entire project?*

e. *What level can the teachers reach in acquiring IAT skills for evaluating their testing items as an outcome of the entire project?*

Hypothesis 6c

Statistically, there are no significant differences between the teachers' acquisition of IAT skills that are associated with any of their researched characteristics.

The purpose of question Q6a is to determine the extent to which the teachers can readily run the CAIAT software. It is one of the most important questions in this project, since obtaining IAT skills is assumed to be the most difficult new concept for teachers. Actually, it is fundamental to the HCD question on construction skill and has a reciprocal[34] effect on teachers' ability about HCD concept skills. The needed information for this purpose is provided by Table 6.13 which presents findings of the observation instrument that was applied during the workshop (while the teachers were using the CAIAT software). The observation aims were to have a hands-on reflection about the level the trainees reached for using the CAIAT software and their level of comprehension of IAT concepts that underlies its outcomes; which could also contribute to answering Q6b. The workshop was held just after the training course, thus this observation's findings are considered to be an immediate indicator of the functionality of training in terms of what it was intended to measure[35].

Table 6.13 illustrates the workshop observation's findings in two sections entitled respectively: “Running CAIAT Software” and “IAT skills”. Percentages of the first section of Table 6.13 range between 90.6% and 99%. This indicates a high level of acquisition of the required skills for running the CAIAT software as an outcome of the short-term training for the project, which answers Q6a. In terms of the IAT skills, which is the concern of Q6b, the second section of this table deals with this area. These data are according to the research assistants' observations and interviews. They were sharing the teachers a round-table panel session. In this method, each of the teachers explained her interpretations/comments about the output report she obtained by the CAIAT, which revealed to the research assistant what level of applying IAT skills the teacher had

reached. I Items 8, 9 and 12 are about initial assessment which classifies weak and good items; item 11 provides a thorough diagnosis that goes beyond classification to identifying the probable causes of weakness, which is a basis for what item 10 includes; and item 10 is the final step in utilising IAT, whereby decisions about remedial actions to be undertaken for refining the defective items are identified. The values of all of these items indicating positively that the project succeeded in providing the teachers with these skills. This is triangulated by the pre- and post-test findings. Diagram 6.7 presents the post-test results of IAT skills in a negative skewed curve, and when this curve is compared to that of the pre-test in Diagram 6.6, we note that training in IAT concepts and skills has functioned to shift teachers' abilities towards better levels of the tested skills. Numerically, the average of the pre-test marks is 4.9 while the average of the post-test marks is 8, which indicates quite a good increase and thus confirms what the workshop's observations revealed.

Question Q6c requires the examination of the significance of the statistical differences between teachers' results in the IAT skills' section of the post-test which, according to Table 6.8, have no statistical significant values for any of the teachers' characteristics, except the key stage. This leads to accepting the null hypothesis 6c only for the key stage variable, which means that the differences between the teachers' results in the IAT section of the post-test could be interpreted only by the key stage in favour of those who teach in secondary schools and is not associated with any other teachers' characteristics.

Similarly to Q6a and Q6b, research questions Q6d and Q6e are about the CAIAT and IAT but target the entire project's functionality in these two content sub-objectives. The research attempted to answer these through the CSs. I will present the CSs in a separate section below, but for now, a glance at Table 6.14, which summarises that section, shows that in the IAT findings row of the table, CSs revealed the good level of ability which the teachers have achieved both in the CAIAT and IAT areas. In addition, teachers' comments and the benefits of the lessons that they articulated at the end of the CSs reflect the good level they reached in terms of running the CAIAT and employing IAT for improving their test construction practice.

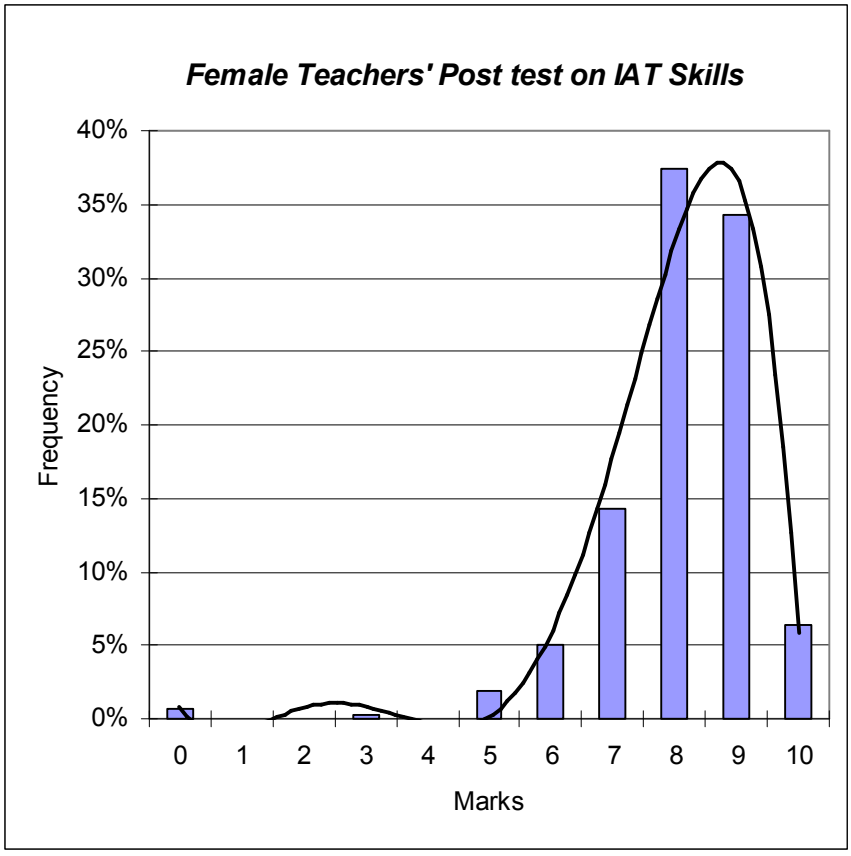
In summary to answer Q6, the project was effective in providing the teachers with skills in the CAIAT and IAT for both the short and long-term outcomes. However, their short-term acquisition of these skills is not associated with any of the teachers'

researched characteristics except for the key stage where more acquisition has taken place by secondary school teachers.

Table 6.13: Summarized findings of the workshops' observation instrument (IAT dimension)

	Observational Guides	Yes	No
Skills of running/implementing the CAIAT Software			
1	Do they have the ability to run the software?	99%	1%
2	Do they know how to enter the initial specification of the test?	94.3%	5.7%
3	Do they know the meanings of its menu components?	93.9%	6.1%
4	Do they know how to enter marks, using the software?	93.2%	6.8%
5	Do they read through the printed reports easily?	92.7%	8.3%
6	Can they review entered information easily?	92.5%	7.5%
7	Can they edit any data entry errors easily?	90.6%	9.4%
Skills of utilising IAT			
8	Do they possess enough ability to single out the good test items from their analyses results?	95.8%	4.2%
9	Do they possess enough ability to single out the suspected weak test items from their analyses results?	93.9%	6.1%
10	Do they have the ability to refine those weak items, using the criteria of optimum test questions?	93%	7%
11	Do they have enough ability to find out what is the probable cause of the suspected weakness of an item?	92.06%	7.94%
12	Do they find any difficulty in comparing two coefficients to each other to gain one judgement? (For example, difficulty coefficient versus discrimination index).	26.2%	73.8%

Diagram 6.7: Distribution of participants' post-test results in IAT skills



Marks	Frequency
0	1%
1	0%
2	0%
3	0%
4	0%
5	2%
6	5%
7	14%
8	37%
9	34%
10	6%

6.3.1.2.3 The Case Study Findings

The aim of the case studies is to discover the extent to which the teachers were able to apply the concepts and skills of the project; therefore, these case studies provide a hands-on measurement of the functionality of what the new intervention implies. Unlike observation, which measures effectiveness as an immediate outcome of the training, CSs measure the outcome of the entire project. They also work in triangulating question Q4c. The effectiveness of the acquisition of IAT skills, which Q6e includes, is targeted mainly by the CSs because looking at the level that a teacher could reach in making decisions about something needs direct discussion and dynamic interaction rather than mere observations and/or content analysis of her/his paperwork. The CSs' findings are presented in a detailed narrative description in Appendix 10; however, I will present from that narrative selected parts that explain the core findings that relate to the research questions. In addition, a summary of the main findings is shown in Table 6.14.

Table 6.14: Summary of the case study findings

Item	1st case	2nd case	3rd case	4th case	5th case
Key stage	Intermediate	Intermediate	Secondary	Intermediate	Secondary
Subject of specialisation	Chemistry	Physics	Physics	Chemistry	Biology
Years of experience	13 years	6 years	6 years	13 years	6 years
Findings Summary for Effectiveness Dimension					
Level of running the CAIAT	*Very High*	*Low to Moderate*	*Moderate*	*High*	*Not available*
Level of IAT interpretations	*High*	*High*	*Very High*	*Low to Moderate*	*Low to Moderate*

First Case

Use of the CAIAT Software

During this teacher's use of the CAIAT, she was quick and accurate in following the steps of entering the initial data such as adding a class, subject or pupils' names. She raised some new issues in using the software, such as copying and pasting data from one location to another. This indicates her mastery of using the package, which most probably, occurred as a result of her intensive use after the training period. All of these aspects are positive indicators of what Q6d seek to explore.

She used the 'repeated mark' advanced data entry function of the software. Furthermore, she provided a list of suggestions to improve the software. Consequently, I would classify her as an optimal user of the software, who has interacted with its purpose actively enough to give a good reliable judgement and reflection about its usage and functionality both from her verbal responses and her observed behaviour.

Applying IAT

The teacher was able to read the results of the CAIAT software outputs accurately and succeeded in correlating difficulty and discrimination parameters. On the other hand, she succeeded in identifying the good items that gained optimal parameters. More observations are elaborated in Appendix 10. At the end of this subsection, I will provide an overall comment for the shared observed qualities of the five cases.

Second Case

Use of the CAIAT Software

This teacher's use of the software was generally moderate. Although she is not a professional user and had faced some minor problems, she was able to continue and enter data. The program menu and the meaning of its contents were clear to her, and her speed in entering data was moderate with few errors. She was able to open a class file but needed more than one attempt to succeed; she then concluded by adding the initial data of the pupils and items successfully. However, there were some times when she forgot to enter the item standard mark or to define its type. This is because she did not attempt to use the software more than once during the contingent stage of the project, which was after the training session. Finally, although she was not classified as a professional user of this software, she succeeded in obtaining a report of the item analysis and was able to make use of the software on her own. All of these aspects are positive indicators of what Q6d seek to explore

Applying IAT

She was able to identify weak and strong items and highlight some recommendation for future implementation. She also connected what she found to her instruction practice, which should provide her with good feedback and on-going improvement in terms of professional development. This should give the answer to research question Q6e.

Third Case

Use of the CAIAT Software

During the interview the teacher was somewhat confused (or maybe frightened about the situation), which affected the level she performed at while running the software. She needed many attempts before she was able to initialize a new class and subject. However, later this apparent nervousness disappeared and she was able to perform at a higher level.

Her use of the program was generally at a moderate level. She is not a professional user and she faced some problems using the software. However, she was able to continue and enter data. She entered a set of pupils and the initial data items but her work was not very accurate: she missed entering the item standard mark during the “add new item” function. However, she discovered her problem and was able to modify

the items by means of the “edit item data” function. Moreover, during data entry, she forgot to enter an item mark; but later she noticed that the sum of all the marks was not the same on that pupil's paper. Consequently, she discovered the missed mark and entered it. Her overall data entry speed was moderate, with the exception of the multiple-choice items, which was fast due to the similarity of alternatives that has been chosen by most of the pupils. In general, I consider this case as an example for most teachers' first attempts at using the software highlighting that she succeeded in acquiring what the Q6d seeks to explore.

Applying IAT

She was able to read the analysis report and correlate between the two parameter values commenting on the decrease of some of the item discrimination statistically as a result of their high difficulty value(i.e. being very easy ones). She noticed an item of a negative discrimination value but reviewed the data entry process and found some errors. She classified three items as good items since their difficulty and discrimination were 75% and 50%, 63% and 25%, and 56% and 38% respectively. I read through these to judge her conclusions and found it valid. The first two items require reasoning and their content is not a very common idea which pupils usually keep in focus; both aspects make them not easy items as her justification had indicated. The third item is a problem-solving question, which can be difficult as well.

She criticised an item albeit its parameters were 63% and 50% because she anticipated that the content of this item should be very easy for pupils and the difficulty value should have been more than 63%. However, she justified this result by the fact that the content comes in two different locations within the textbook, which might have distracted the lower-achieving pupils. In my opinion, this justification is reasonable and can be true in some sense. She reflected that the pupils should be alerted to this issue in the test feedback session.

She commented on 100% difficulty and 0% discrimination as being for weak items because most of them were at recall level. However, there were some problem-solving questions, but she indicated that they had been explained many times to the pupils, which could have made the pupils' cognition of this curricular content similar to that of recall level. She commented successfully that some alternatives should be revised to be more efficient. One item of a distinguished discrimination of 0% whereas difficulty was 75%, revealed a paradox. She went back to the pupils’ answer-sheets and

found a reasonable cause. She found that the efficiency coefficient of one alternative was 0% and by a relevant analysis she indicated that its content could be true in some sense whilst it is supposed to be false as a distractor.

Fourth Case

Use of the CAIAT Software

This teacher's use of the program was generally good. She is not a professional user and faced a problem during the setup where she had to try to use the software on more than one PC but at the end she was able to succeed. During the entry of the items' marks she was slow but in the end she was able to analyse and obtain the results report successfully which answers Q6d positively.

Applying IAT

Her technical reading of the analysis results was not high because she classified the items according to their difficulty coefficient without looking at the discrimination. Her justification is that she did not understand the discrimination parameter. Nevertheless, her explanations of the results are generally satisfactory since she was able to justify her opinions by the standards for question construction. She also was keen to look through pupils' answer-sheets for further understanding. I will elaborate on some of her interesting explanations.

She considered the difficulty value of 71% for one item as above her anticipated value because the HCD skills for answering require many steps. However, she justified this high value (easy item) by the way she had presented this subject during her lessons where she depended on a story-telling technique. This could have facilitated the pupils' retention of related facts. I think that this is one of the most important aspects of professional reflection that teachers can gain from this semi-AR practice. Adversely, she denoted that another item's difficulty value of 56% not high enough for what she was anticipating according to the overall easy level of the content. She suspects that the linguistic difficulty of the term 'taxis' is a probable cause. I think this is not necessarily the reason because the terminology as part of the curriculum should be clear to the pupils. However, the teacher's expectations might have been too optimistic compared to her less qualitative instruction for this part of the curriculum which could be a cause. If this is true, I think that as her future trials and analysis focus on the reason she suspects, she will find that she still has the same issue. She is most likely at that time to change

her mind about this justification and look from another angle which relates to her instructional practice rather than from the specific piece of reason which she outlined. The educational supervisor's role in this sort of practice should help in promoting better scaffolding for the teachers' learning.

Her justification about the reasons that underlie the weakness of some of the questions does not reflect a good level of acquisition of HCD question construction skills, which are targeted by research question Q4c. This is because she focused on connecting between the levels of cognition and the types of questions as a framework for judging the items' weakness, and on improvement suggestions, which are the easiest and most fundamental approach for improving HCD questions. Looking at the language of the question is more important in this regard and reflects better ability and experience which she did not give a sound attention.

Fifth Case

Use of the CAIAT Software

Unfortunately, this teacher's use of the CAIAT software has not been reported because she thought that the CS was concerned with the IAT sessions only thus had her use of the CAIAT in the absence of the research assistant.

Applying IAT

She was able to read the results of the CAIAT software outputs accurately identifying good and weak items upon their parameters. She justified the higher results (easy items) by the frequent training she delivered to her pupils for the skill related to those questions and by the illustrative drawing that she used as a teaching aid which contributed in the long lasting retention of the related information.

The presentation in Appendix 10 for the five cases' interpretations of the results they found by the CAIAT indicates that they all highlighted successfully a number of items with a relevant use of IAT, showing a reasonable level of understanding of what the project's training aims are in terms of IAT skills. This should answer the research question Q6e positively. Furthermore, their justifications about the reasons that underlie some weak questions reflect the extent to which they have acquired the skills for writing HCD questions which is targeted by research question Q4c (except for the

fourth and fifth cases). They commented using different conceptual frameworks according to the following in relation to each case:

1. The first case mentioned Bloom taxonomy, wording errors in constructing multiple-choice items and distractor construction.
2. The second case mentioned Bloom taxonomy and distractor construction.
3. The third case connected her opinion to the curriculum, to the distractors' efficiency, and to Bloom taxonomy.

Comment on All Cases

The CSs findings, as shown by Table 6.14, reveal that the teachers studied have acquired the required skills of the CAIAT software and IAT with fairly good levels of attainment. To find an average of these, I will quantify these out of a hundred on the following basis: 25 for low, 50 for moderate, 37.5 for their average, 90 for high and 95 for very high. In this case, the average of running the CAIAT level is 68, which is moderate, and the average of the IAT skills level is 70 which is moderate to high. The qualitative investigation of the cases studied indicates a similar result. To elaborate, the teacher in the first case showed a distinct high level of using the CAIAT software since she used it many times during the contingent stage, was quick in using it, efficient in choosing what function or operation to perform for every step and suggested many ideas for improving the CAIAT that reflect the extent to which she was enthusiastic about the software and interacted with it positively. The teacher in the fourth case showed a similar level with some differences in terms of the speed and intensive prior use of the package. However, she was able to use the software without significant errors or problems such as those that occurred in the second and third cases, thus she was considered a good user of the software. The teachers in the other two cases were able to overcome the problems or obstacles that they faced, which indicates a good possibility for the software to be learnt simply on a trial and error basis, hence the four cases reveal a general success of the entire project's effectiveness in using the CAIAT.

As for IAT skills, the teachers in all cases except the fourth succeeded in correlating difficulty and discrimination parameters, and all provided an adequate analysis that showed the reasons for their findings in the light of the standards for test item construction, especially wording, as reported for the first and the second cases. The teachers in the second, the third and the fourth cases linked their analysis of some of the items to instruction. The teacher in the third case succeeded in connecting between

some questioned data by reviewing the data entry and correcting the errors that she found. In the third and the fourth cases, the teachers connected some issues to the curriculum and adopted the practice of reviewing pupils' answer-sheets. As I mentioned earlier, this successes of the cases triangulate the positive result about the CAIAT and IAT which was found through the workshop observations and pre/post-tests.

6.3.1.3 Research Questions about Functionality of Training

6.3.1.3.1 Research Question 7a

Q7. What level of ability can the researched teachers acquire from the training of the project in general?

a. *Is there any improvement when comparing the overall pre- and post-test results?*

b. *Can this improvement (if any) be interpreted by the training factor?*

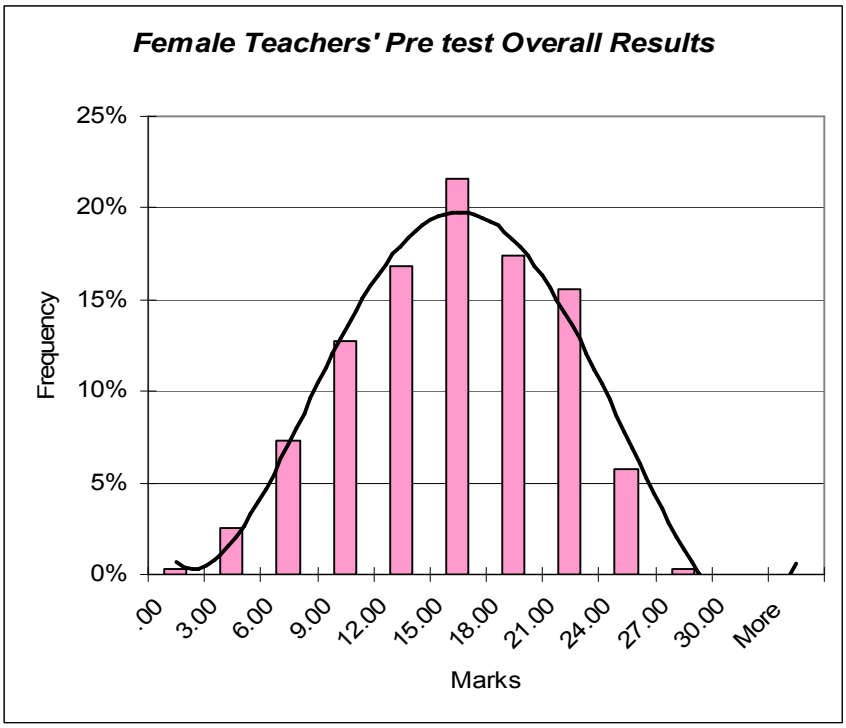
Hypothesis 7b

Statistically, there are no significant differences between the teachers' pre-test results and post-test results which are associated with the training of the project .

Overall Results

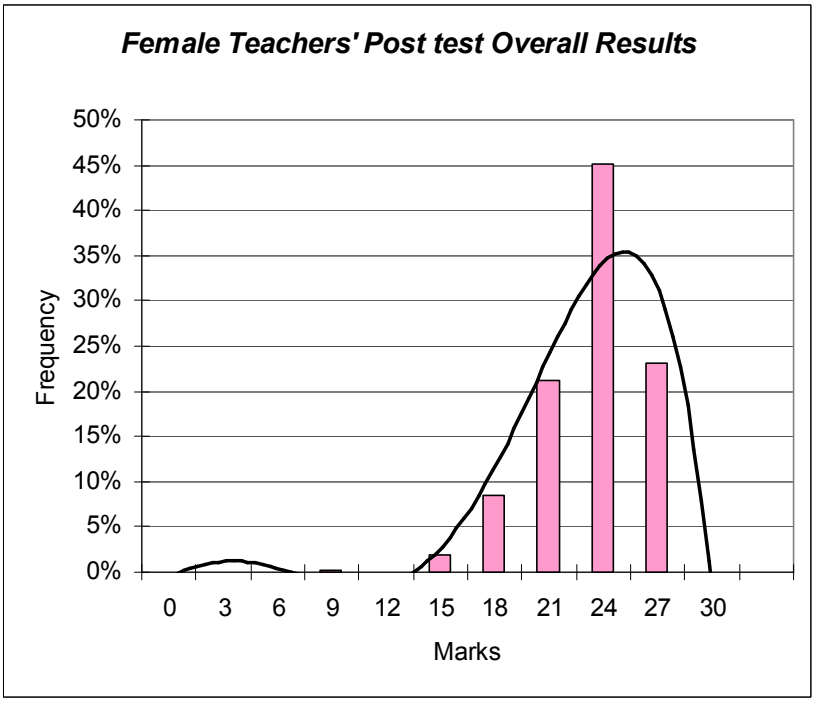
The pre and post-tests consist of sections that each of which reveal the functionality of the training in the dimensions related to that section (HCD or IAT). The pre-test overall results represent all of the sections results aggregated together to reveal functionality of the training as a whole. This type of findings appears in Diagram 6.8 for the pre-test on a normal curve, which is the usual distribution of academic achievement outcomes. The post-test overall result appears in Diagram 6.9 showing a negatively skewed curve, which represents a high level of ability for the individuals tested. The average of pre- and post-test overall marks has increased from 17.8 to 26.3, which indicates the good level of benefit which the participant teachers have obtained from the training of the project as an overall result.

Diagram 6.8: Distribution of participants' pre-test overall results



Marks	Frequency
0	0%
3	3%
9	7%
12	13%
15	17%
18	22%
21	17%
24	16%
27	6%
30	0%

Diagram 6.9: Distribution of participants' post-test overall results



Marks	Frequency
0	0%
3	0%
9	0%
12	0%
15	0%
18	2%
21	9%
24	21%
27	45%
30	23%

Table 6.15 shows that T value has statistical significance, thus the null hypothesis 7b is to be rejected and differences between the pre- and post-test overall results could be interpreted by the training, which indicates its importance and its good level of functionality. Some comments from participant teachers as reported by the

research assistants expressed that their benefit from training was high and rich. One female research assistant reported that "a number of teachers have shown greater interest in the project than some educational supervisors have." Another comment mentioned that many of the teachers were encouraged by the fact that some technological aspect is involved in this sort of self-learning practice, as it creates a challenge and curiosity for discovering what their 'numerical picture' could be in terms of the quality of their tests[36].

Table 6.15: Paired samples T-test for the teachers' overall results of the pre- and post-tests

Test	Average	Number of Cases	Standard Deviation	Degree of Freedom	T value	Significance
Pre-test	4.6091	301	1.77195	300	-23.345	0.001 *
Post-test	7.4779	301	0.91887			

Partial Results

To consolidate the above overall finding, I will show the extent to which training has impacted different sections of the project; namely: acquisition of HCD concepts, skills in writing HCD questions and acquisition of IAT concepts and skills. Table 6.16, which compares the pre- and post-test results of the different sections of training, reveals that differences between teachers' results are statistically significant for all of these sections, which indicates the value of the project's training for each section confirming the positive functionality found for the overall results of the training.

To summarise in answering Q7, the project training has raised the participant teachers' ability to a better level. Furthermore, it is found statistically significant that the source of difference between before- and after-training data is the training. This source is found to come from all sections of training, which indicates the value of each section as well as the value of the training course as a whole.

Table 6.16: Paired samples T-tests for teachers' pre- and post-tests results of the different sections of the training

Content Objective	Section of Training	Test	Average	Number of Cases	Standard Deviation	Degrees of Freedom	T value	Significance
HCD	Knowledge about HCD concepts	Pre-test	**4.4381**	**315**	**2.99976**	**314**	**-18.427**	**0.001 ***
		Post-test	**7.9206**	**315**	**1.24027**			
	Skills in writing HCD questions	Pre-test	**4.5993**	**302**	**2.01821**	**301**	**-11.158**	**0.001 ***
		Post-test	**6.3626**	**302**	**1.81773**			
IAT	Acquisition of IAT skills	Pre-test	**4.9460**	**315**	**2.73226**	**314**	**-18.910**	**0.001 ***
		Post-test	**8.1270**	**315**	**1.17690**			

6.3.1.4 Synopsis of Effectiveness Dimension

- The participating teachers, generally, have a limited previous background of HCD concepts and skills; however, this is at a higher level for those educationally qualified; those who have previous training in test construction or IAT skills; those who teach in a secondary school; those from Biology or Physics specialisations for the HCD concepts; and those from Chemistry or Biology specialisations for the HCD question construction skills.
- Training the teachers in HCD concepts and skills is found to be functioning and independent from all teachers' researched characteristics except that more acquisition of HCD question construction skills has taken place with secondary school teachers.
- Observation reports indicate that the teachers have shown good abilities in acquiring skills for asking valid HCD questions during instruction and within tests as a long-term outcome of the entire project. Statistically, the training effectiveness in improving the teachers' abilities was found significant, which is triangulated by the qualitative data of the interviews.

- The teachers' previous background in IAT concepts and skills is limited and has no association with any of the teachers' characteristics except for those having training in test construction who were found to be better in this regard.
- There is an acquisition of the required skills for running the CAIAT software and IAT skills for short and long-term outcomes. In terms of short-term outcomes, secondary school teachers tend to acquire this skill much better than others and no other teachers' researched characteristics were found to be associated.
- The project was found to be statistically significant in raising the participant teachers' ability to a better level for all skills compared to their previous levels. Furthermore, all sections of training were found to be valuable in this respect.

6.3.2 Adoption Dimension

This dimension aims to explore the extent to which the teachers were enthusiastic in applying what the project implies, especially for its two main concepts, HCD and IAT. In practice, the former concept is to be embodied in their use of HCD instructional objectives and their HCD assessment questions, whether during lessons or within their TA tests. The latter is to be embodied in their interest in using the CAIAT software; which also works as an indication of their interest in its underpinning concept; that is, IAT. Table 6.17 includes the data of the questionnaire responses that will be used for answering what the adoption dimension research questions seek to explore (Qs 8, 9 and 10). Table 6.18 for open-responses is utilised to cross-validate what is found by Table 6.17. This sort of treatment is because adoption is contingent on the one hand and difficult to be measured on the other, so this entails more than one approach being followed in order to make sense of the research data obtained. Table 6.19 presents the results of hypotheses testing for examining whether the research variables have any statistically significant impact on the teachers' adoption of them.

6.3.2.1 Adopting HCD

6.3.2.2.1 Research Question Q8

Q8. To what extent will the researched teachers adopt what the project implies about HCD instructional objectives?

a. To what extent do the teachers think that the project has an impact of encouraging them to increase their use of HCD instructional objectives?

b. *With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).*

Hypothesis 8b

Statistically, there are no significant differences between the teachers' use of HCD instructional objectives that are associated with any of their researched characteristics.

To answer Q8, Table 6.17 shows the teachers' perspective about their levels of interaction with the project. They reported that the training course had a positive effect on their use of HCD instructional objectives. 94.9% of the teachers (according to item 3 of Table 6.17) agreed that this aspect had increased as a result of their attendance at the course. Of the 94.9% teachers, 50.5% estimated the ratio of their increase in skill level as highly as from 70% to 90% while 34.9% estimated the ratio of increase to be no less than 50%. Totalling both findings yields that 85.4% of the 94.9% of teachers, which is 81% of the whole sample, estimated that there was an increase in their use of HCD instructional objectives of a percentage not less than 50% which justifies the consideration that the project's related objective has been accomplished successfully. The teachers' researched variables were found not to have an impact because none of the T and F values are statistically significant according to Table 6.19.

The answer for Q8 is that the teachers' adoption of what the project implies about HCD instructional objectives is positive and is free from any of the teachers' researched characteristics.

6.3.2.2.2 Research Question Q9

Q9. To what extent will the researched teachers adopt what the project implies about asking/writing HCD questions?

a. *To what extent do the teachers think that the project has an impact on encouraging them to increase their questions at the HCD level?*

b. *With which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).*

Hypothesis 9b

Statistically there are no significant differences between the teachers' practices of asking questions at HCD level that are associated with any of their researched characteristics.

Because I am going to utilise Table 6.18 in answering this research question, besides Table 6.17, it is necessary to highlight that because it includes free response

findings, the presented percentages should not be looked at in the same way as one looks at the percentages in a closed questionnaire. In this sense, the significance of a percentage of 50%, for example, is not that half of the respondents are those who support the related issue but that there are 'many' individuals who have been articulating their need for or opinion on that issue from the beginning. This means that they give this issue a priority or importance rather than mere agreement. At the same time there is an open likelihood that more individuals might have had the same interest but they did not articulate it. References in research methodology that I have reviewed do not mention how percentages of open-ended coded findings can be interpreted. Cohen et al. (2000: 255) highlighted the danger of open-ended data handling when some tend to aggregate these data into numbers and treat these in the same way as quantitative data. They believe that this danger comes from the violation of borrowing from the quantitative paradigm to use the data in a different, qualitative paradigm. I would suggest that percentages of more than 10% in this respect are accepted indicators, more than 25% would be strong ones and more than 50% should be highly significant[37].

Table 6.18 classifies the teachers' open responses into five main categories: benefits, obstacles and difficulties, suggestions, feelings, and reflection on pupils. When discussing this table's findings, I will focus on the items of high percentages and those tackling closely related aspects of the project. Generally though, it could be noted that most of the teachers' responses tackled the CAIAT dimension and in the second order comes the HCD dimension; fortunately, this meets this research's intention which considered IAT the prime input dimension. The items that inform HCD input dimension are 3, 4 and 6 and the items for IAT are 1, 2, 8, 10, 11-15, 20-22, 29, and 30.

For answering Q9a, Table 6.17 reveals that the project had a positive effect on the teachers' practical interest in using HCD questions during instruction and within their tests. This is stated by 95.8% of the teachers, according to item 4 of Table 6.17. Out of these, 58.5% think that the percentage of increase level in this practice is in the range of 70% to 90% while 32.4% estimate it to be no less than 50%. This yields that 90.9% of the sub-sample, which is 87% of the whole sample, estimate this increase as greater than 50%, which justifies addressing the success of the project in eliciting the researched teachers' use of HCD questions. This result is triangulated by the findings of the open-ended questionnaire where items 3, 4 and 6 of Table 6.18 show that a high percentage of the teachers' responses agree on the benefit of the project in increasing

their ability to write good HCD questions. Item 7 reveals that this would accordingly increase their teaching efficacy.

Table 6.19 shows that none of the related T and F values have statistical significance except for the fact that T value = 0.014 since it is significant below 0.05. This leads to accepting the null hypothesis 10b for all factors except educational qualification. The differences between teachers in their adoption of using questions of HCD level during instruction are interpreted only by the factor of educational qualification in favour of those educationally not qualified[38]. In other words, the teachers who graduated from colleges other than Colleges of Education are more likely to adopt HCD questions during instruction or within tests as a response to the project.

The answer for Q9 is that the researched teachers' adoption of what the project implies about HCD questions is positive and is free from any teachers' researched characteristics except educational qualification where the teachers from colleges other than Colleges of Education are more likely to adopt HCD questions.

6.3.2.2 Adopting IAT

Q10. To what extent will the researched teachers adopt what the project implies about the CAIAT software and IAT skills?

a. *To what extent will the teachers use their own initiative in applying the CAIAT software?*

b. *To what extent will the teachers' professional development be elicited by the CAIAT software?*

c. *In terms of applying the CAIAT software on the teachers' own initiative, with which teachers' researched characteristic(s) are statistical differences between the teachers associated? (If any).*

Hypothesis 10c

Statistically, there are no significant differences between the teachers' use of the CAIAT software on their own initiative, by a self-learning basis, that are associated with any of their researched characteristics.

Firstly, I have to clarify the difference between Q10a and Q10b, where the former tackles an initial level of interest that is represented by installing the software and using it at least once. Items 1 and 2 of Table 6.17 serve this aim. The latter aims to find out whether there is a further level of interest that is represented by a full

employment of the software and applying IAT concepts and considerations to reading through the report results produced by the CAIAT, and hence utilising this reading to identify how to improve that test. Item 5 of Table 6.17 serves this aim.

Items 1 and 2 of Table 6.17 reflect a positive answer about whether the teachers have shown any initial interest in the CAIAT. The majority of the teachers (88.9%) have a computer at home and the majority of these users (81.3%) have installed the CAIAT software on their machines, which indicates that they are interested in the software's function. Also, 81.4% of the teachers tried to use the software on their own, at least once, in order to discover the quality of their testing items. Looking further at the statistics, one finds that those who tried to use the software three times or more are 30% out of the group of 81.4% (by adding the last two percentages 11% and 19%). In fact, this means that 24% of the whole sample tried it three times or more. Having almost a quarter of the entire sample at this fair level of initial interest is a further indicator of the willingness to adopt the software. These findings should answer what Q10a asks about and indicate that the sample teachers were, initially, interested in the software's function.

For Q10b, which seeks to explore the teachers' long-term adoption, item 5 reveals that 79.2% of the teachers have tried out the CAIAT software on a self-taught basis. Adding up the two percentages 46.2% and 33.3% yields that the majority of this group (almost 80%), who used the CAIAT on their own, estimate their success as no less than 60%. This answers Q10b positively. In addition, the open-ended questionnaire/report shown by Table 6.18 cross-validates this finding where items 1 and 2 of Table 6.18 express, through the highest rate of responses (75% and 72.5% respectively), that the participant teachers articulated their acceptance of the main concept of the project. Items 8, 29 and 30 indicate, through a moderate level of response, that they acknowledge the CAIAT easiness value in increasing test validity by doing item analysis.

For Q10c, Table 6.19 shows that the related T and F value[39] have no statistical significance; which leads to accepting the null hypothesis 10c. This means that the differences between individuals in their initial use of the CAIAT software on a self-learning basis are not interpreted by any of the researched factors.

To summarise in answering Q10, the teachers have adopted the CAIAT software on their own initiative, and their long-term PD was elicited by the CAIAT software and is found free from any of the teachers' researched characteristics.

Table 6.17: Questionnaire for the contingent application stage

Question	*Yes*	*No*	*Missing*
1. Do you have a personal computer at home?	**88.9%**	**11.1%**	
Have you installed the CAIAT software on your own PC?	81.3%	18.7%	
2. If you have installed the CAIAT software on a PC, have you tried it out for discovering the quality of your test questions? (Regardless of whether you succeeded in this/these attempt(s) or not)	**81.4%**	**16.6%**	**1.8%**
If 'Yes', how many times:			
Once	44%		
Twice	26%		
Three times	11%		
4 – 10 times	19%		
If 'No', Reasons: *Note that ratios here do not add up to 100% since they may overlap.*			
Reason: Difficulty in installation	44%		
Reason: Limitation of time	11%		
Reason: No experience in using a PC	17%		
Reason: Work overload	5.5%		
Reason: Reason not mentioned	5.5%		
3. After attending the training course, has your use of HCD behavioural objectives for instruction increased?	**94.9%**	**5.1%**	**-**
If 'Yes': **How much do you estimate the ratio of increase to be:**			
from 70% - 100%	50.5%		
from 50% - 69%	34.9%		
less than 50%	14.6%		
If 'No': Reasons: *Note that ratios here do not add up to 100% since they may overlap.*			
a. Not interested in this issue.	0%		
b. Do not have much time.	9.1%		
c. This adds to my workload.	18.2%		
d. The school does not appreciate such improvement in my ability.	0%		
e. I feel afraid that I might make some scientific errors if I tackled HCD, thus I tend to be limited to the lower cognitive level.	18.2%		
f. I think that teaching science/physics should not go as far as the level of higher cognitive demand.	0%		
g. I did not understand how I can apply HCD concept implications in the real world.	0%		
h. Although I understand what I have learnt on this course, I think I need more time until I understand it thoroughly and am able to apply it.	27.3%		
i. Other reasons: Because I have many subjects to teach.	36.4%		
4. After attending the training course, have your questions at the level of HCD, either during instruction, in worksheets, in home work or in	**95.8%**	**3.2%**	**1%**

your tests, increased?			
If 'Yes': How much do you estimate the ratio of increase to be**:**			
from 70% -100%	58.5%		
from 50% - 69%	32.4%		
less than 50%	9.1%		
If 'No': Reasons: *Note that ratios here do not add up to 100% since they may overlap.*			
a. Not interested in this issue.	0%		
b. Do not have much time.	28.6%		
c. This adds to my workload.	0%		
d. The school does not appreciate such improvement in my ability.	14.3%		
e. I feel afraid that I might make some scientific errors if I tackled HCD, thus I tend to be limited to the lower cognitive level.	14.3%		
f. I think that teaching science/physics should not go as far as the level of higher cognitive demand.	0%		
g. I did not understand how I can apply HCD concept implications in the real world.	0%		
h. Although I understand what I have learnt on this course, I think I need more time until I understand it thoroughly and am able to apply it.	18.6%		
i. Other reasons: Because I have many subjects to teach.	28.6%		
5. After attending the training course of HCD-CAIAT project, have you tried to use what you learnt about the CAIAT software on a self-learning basis?	**79.2%**	**14.3%**	**6.5%**
If 'Yes': How much do you estimate the ratio of your success in this respect to be?			
from 80% - 100%	46.2%		
from 60% - 79%	33.3%		
less than 60%	20.5%		
If 'No': Reasons: *Note that ratios here do not add up to 100% since they may overlap.*			
a. Not interested in this issue.	6.5%		
b. Do not have much time.	51.6%		
c. I am very confident that my ability in writing test questions does not need any improvement.	16.1%		
d. The school does not appreciate such improvement in my ability.	9.7%		
e. I did not understand what the project is all about.	6.5%		
f. This adds to my workload.	38.7%		
g. I did not feel confident that I can apply what I learnt in this course.	9.7%		
h. I need some time until I understand thoroughly what I have learnt.	29%		
i. I do not agree that this is a way to improve test item construction.	25.8%		
j. Other reasons: 1) Being new to using computers. 2) Limitation of time because I have many subjects to teach.	6.5%		

Table 6.18: Findings of the open-ended questionnaire

Category	Response	F	P
Benefits?	1. Providing teachers with the ability to judge their questions in a well-structured way.	30	**75%**
	2. Providing teachers with skills of using computers through a new software package.	29	**72.5%**
	3. Developing teacher's skills in HCD question construction.	26	**65%**
	4. Developing teacher's skills in ways of asking HCD classroom questions.	19	**47.5%**
	5. Exchanging experiences amongst participating teachers and between teachers and educational supervisors.	16	**40%**
	6. Providing teachers with skills in writing HCD behavioural instructional objectives.	14	**35%**
	7. Increasing instruction efficacy.	14	**35%**
	8. Helping the teacher to know the level of difficulty or easiness of her questions, so as to consider this in future tests.	7	**17.5%**
	9. It helps to create an items bank.	3	**7.5%**
Obstacles and Difficulties?	10. Difficulty of installing the software on some PCs.	21	**52.5%**
	11. Using this software adds to the workload of teachers who have timetables with a high rate of lessons besides having multiple subjects	19	**47%**
	12. If there is a large number of pupils and items then the data entry would take a lot of time.	18	**45%**
	13. There are not enough extra PCs in the school and the principal sometimes refuses to allow the teachers to use the school's main computer for the CAIAT project because she is afraid other software packages or school data might become corrupted.	14	**35%**
	14. Assigning other tasks to the teacher such as extra-curricular activities adds to their workload, besides using this software.	14	**35%**
	15. It is difficult to apply when there is a huge number of pupils.	11	**27.5%**
	16. The software or its data cannot be copied onto a USB or any other external means of data storage.	8	**20%**
	17. The software may stop during work or may lose some data after switching off the machine.	6	**15%**
	18. The software does not accept decimals.	6	**15%**
	19. Sometimes I get an error message from the software when I ask for analysis report or maybe I get a blank sheet.	5	**12.5%**
	20. Pupils and their parents refuse to let teachers ask questions at HCD level.	4	**10%**
	21. There are a number of teachers about to retire therefore it is no use training them to use such new technology.	3	**7.5%**
	22. Principals do not have a good idea about the importance of this method which could have encouraged them to cooperate in its employment.	2	**5%**
	23. If there is an error in entering the marks then no error message appears instantly but the software will hang at the "running the analysis" function.	2	**5%**

Suggestions	24. The project time needs to be extended so teachers could apply its inclusions much better	12	**30%**
	25. There is a need for further training on constructing HCD questions.	8	**20%**
	26. The software does not use copy and paste facilities.	3	**7.5%**
	27. Results reports do not show names of subject, class or teacher.	1	**2.5%**
Feelings	28. Indicators cannot be trusted since they depend on pupils' levels, which is changeable from one group to another.	10	**25%**
	29. Ease of using the software because its menus are straightforward and in Arabic.	4	**10%**
	30. The software provides a fast method of analysing test results.	4	**10%**
Reflection on pupils	31. Increasing pupils' academic levels and developing their thinking skill.	23	**57%**
	32. Increasing pupils' achievement levels.	9	**22.5%**
	33. Creating an open educational atmosphere through open discussion between the teacher and her pupils, which reinforces a self-learning attitude.	9	**22.5%**
	34. Discovering pupils' abilities and giving them the opportunity to express themselves through their opinions.	7	**17.5%**
	35. Connecting subject content with pupils' lives and the possibility of explaining surrounding phenomena.	2	**5%**
	36. Enabling pupils' knowledge to grow.	1	**2.5%**

Table 6.19: Independent samples test (T values) and One-way ANOVA test (F values), Summaries of *Adoption* Dimension[40]

Targeting differences between individuals upon their:	**HCD**		**IAT**
	Use of HCD Objectives	Use of HCD Questions	Use of The CAIAT Software
	T Values' Significance		
Educational qualification	**0.466**	**0.014 ***	**0.823**
Training in test construction	**0.951**	**1.000**	**0.847**
Training in IAT	**0.835**	**0.399**	**0.051**
Key stage (intermediate/secondary)	**0.204**	**0.492**	**0.244**
Do you know how to use computers?	**0.290**	**0.897**	**0.774**
Do you have a PC at home?	**0.446**	**0.764**	**0.706**
Can you use Excel software?	**0.617**	**0.893**	**0.872**
Can you use Access software?	**0.381**	**0.514**	**0.819**
	F Values' Significance		
Level of graduation	**0.326**	**0.055**	**0.428**
Number of years of experience in teaching	**0.564**	**0.755**	**0.231**
Specialization subject (Physics –Chemistry– Biology)	**0.590**	**0.747**	**0.505**

Chapter 7

Discussion and Conclusions

As the preceding chapter has shown, all the findings of this research positively indicate the effectiveness of the project as well as the participating teachers' acceptance of what the project implies. However, there are a number of comments and clarifications in relation to some of these findings which need to be addressed. Therefore, the present chapter will discuss these findings in greater detail, so as to provide a comprehensive understanding of the different issues involved. I begin this chapter with the descriptive statistics that present the teachers' basic data in order to provide a basis for the discussions that follow and to connect these statistics to the main research findings in order to articulate what conclusions these connections reveal. After this, the findings of the two prime dimensions, effectiveness and adoption, are discussed so as to highlight the main points this research has produced, and to connect these with previous studies or related literature. This will be followed by a section that includes both dimensions, effectiveness and adoption, but is dedicated to the findings on the research variables, in order to outline what the impact of these variables reveals in terms of generalisability and in terms of issues that need to be addressed for the organisation/context studied. Since the research variables are many, and hence to facilitate the reading of what Chapter 6 has illustrated in various locations of the text, Table 7.1 summarises these.

At the end I will state concisely the main conclusions of this research and will list recommendations and suggestions that embody the perceived practical benefit of this research effort. As mentioned earlier, there is no similar or closely related work in this area that is grounded in the context of KSA; in other words, there is no frame of reference that I can use for my analysis. Furthermore, similar non-Saudi studies are scarce and lack the interest to tackle teachers' PD factor as this work has done. Therefore, the analytical discussion of the specific contribution which the present research makes to the field of study will be restricted by this fact since it is breaking new ground and establishing a basis for further development in the field of study. However, during my discussions, possible connections between what the findings reveal and the previous research or relevant reviewed literature are raised. Furthermore, the

theoretical model of change (the Personal Power model) adopted for this research intervention is another important aspect that has been given attention in this chapter.

Table 7.1: Summary for associations of teachers' researched characteristics

Research Dimension	Targeting differences between individuals upon their:	Variables that have an association
Effectiveness (Pre-Test)	Background knowledge of HCD Concepts	**Who have the best background of HCD concepts and behavioural objectives?** 1. Teachers who have graduated from a college of education. 2. Teachers who have had prior training in test construction skills. 3. Teachers who have had prior training in IAT skills. 4. Secondary school teachers. 5. Teachers from Biology or Physics specialisations.
	Background in HCD Questions Construction Skills	**Who have the best background in HCD question construction skills?** - Teachers from Chemistry or Biology specialisations.
	Background in IAT Skills	**Who have the best background in IAT skills?** -Teachers who have been trained in test construction.
Effectiveness (Post-Test)	Acquisition of HCD knowledge/concepts	**Who have the best acquisition of HCD concepts and behavioural objectives?** - No difference between the groups
	Acquisition of HCD Question Construction Skills	**Who have the best acquisition of HCD question construction skills?** -Teachers who teach in secondary school.
	Acquisition of IAT skills	**Who have the best acquisition of IAT skills?** -Teachers who teach in secondary school.
Adoption	Adoption of HCD Objectives	**Who have shown the best adoption of HCD behavioural objectives?** - No difference between the groups
	Adoption of HCD Questions	**Who have shown the best adoption of HCD questions during instruction or in TA practices?** -Teachers who have graduated from colleges other than a college of education.
	Adoption of CAIAT/IAT	**Who have shown the best adoption of CAIAT software and employed IAT for improving their tests?** - No difference between the groups

7.1 Descriptive Statistics

Descriptive findings show that almost 75 % of the participants are graduates of colleges of education, which means that almost three quarters of the participants are educationally qualified before service. Levels of graduation are on a normal distribution, but generally tend to be higher levels, as illustrated by Diagram 6.1. According to Table 6.2, the graduation levels of most of the participants (GPA or equivalent) are in the range of good to very good and most of them have spent 10 years or more in teaching; which indicates these are good experienced individuals. All of these indicators reveal that the teachers are of good quality in terms of their pre-service training and other factors. However, findings of research questions 1, 3 and 5 have shown that generally they have limited levels of ability in the two researched content objectives, HCD and IAT. This calls pre-service training in colleges of education into question and also calls for a thorough research that aims at investigating the quality of these programs. This is confirmed by various examples of previous local research, which highlighted the teachers' low level in areas pertaining to assessment (Al-Mazyed, 1975 and Al-Zahranee, 1998) and HCD (Al-Saif, 1982; Ayedh, 1993; and Al-Darweesh, 1999). Al-Zahranee (1998) stressed explicitly that assessment courses at colleges of education should be redesigned to meet the actual field requirements.

Items 2 and 3 of Table 6.3 show that the majority of the sample individuals, almost 89%, had never attended a training course on IAT and quite a high number of them had not attended a training course on test item construction. This indicates that their INSET qualification was poor in this respect. This has been shown in previous related literature: Al-Mazyed (1975), for instance, stressed the lack of professional preparation for the majority of Saudi secondary school science teachers and described most of their practice as being characterised by the traditional, teacher-centred pedagogy; and Al-Saif (1981) indicated that Saudi science teachers at secondary school have insufficient knowledge of learning and methods of teaching and that the majority of them do not consider in-service education important.

Item 4 of Table 6.3 indicates that the teachers need to evaluate their questioning skills with a more precise or ‘diagnostic’ technique, such as CAIAT and its underpinning concepts and skills, so as to discover the nature of the problem that they faced or felt and to reflect on the way in which they could avoid such a problem in future. I think that to have a third of the group demonstrating this need is a reasonable

proportion to raise the researched issue to a level of importance in order to target the hidden difficulty of this sort of assessment quality. In addition, this need cross-validates what was presented earlier about the rationale and significance of the present research.

Item 5 of Table 6.3 shows that a high ratio of participants (77%) claim to know how to use computers. However, Table 6.4 reveals that this skill is concentrated in the ability to use MS Word© software and, to a lesser degree, PowerPoint. Nonetheless, the finding that the vast majority of teachers had this high level of computer use facilitated the project's training task so it could focus on its major purpose without the need to train the teachers in the basics of computers. This was obvious, according to the observations of the research team, where it was not difficult to teach the teachers how to run the CAIAT software. As will be shown later, the most problematic area in this respect was installing the program on the machines not running it, and/or utilizing its various functions and services. Nevertheless, the administration plan of the project allowed for the possibility that some of the teachers might not know how to use computers, thus a special session for training these individuals on fundamental computers skills was held prior to the main course. The remaining ratio of the teachers, who lack the good ability in using computers (23%) have benefited from this course which was carried out by the Technical Support Division of the Informatics and Computers Centre, which belongs to the GDGEA.

7.2 Effectiveness Dimension

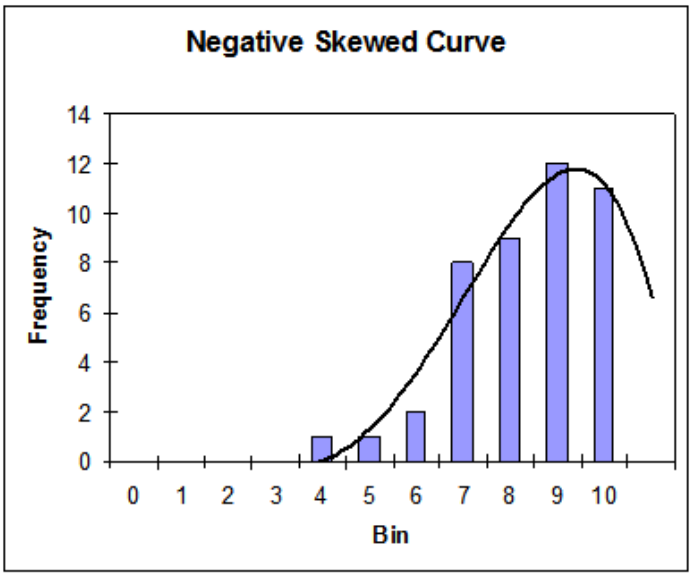
I will now discuss effectiveness of the training course of the project as a whole and then discuss each content objective of the study (HCD and IAT) showing two main points for each dimension: background of the teachers and effectiveness of the training course and of the entire project.

7.2.1 Overall Effectiveness

The pre-test and the initial questionnaire were utilised to discover the teachers' levels before the project. The analyses show that although they have some skills, these are not within the targeted range. Instead of the normal distribution shape in which the resulted diagrams appeared, the suggested form of distribution is assumed to be a negative skewed curve, as the model shows in Diagram 7.1. I consider this an initial indicator for the need for focused training as the design of this project implies.

The post-test diagrams (6.3, 6.5, and 6.7) are almost of a negative skewed shape, which indicates the good level of learning that the teachers achieved from the project training for each of the project's content objectives. Moreover, comparing Diagram 6.9 (the post-test overall results) to Diagram 6.8 (the pre-test overall results), revealed that the teachers achieved good level of attainment from the training as a whole, which has been found to be statistically significant, according to Table 6.15. This is congruent with Hejran's (2002) findings in the KSA context, that the teachers' need for training on assessment comes in first place. This finding is also in line with what Wang et al. (2004) have found: that the teachers who were studied improved in most of the steps of the assessment as a result of using the WATA computer system, which reflected the functionality and hence the importance of their 4-month training. This entails that the WATA project's training package was successful in preparing/empowering the teachers to be able to deal with HCD and IAT.

Diagram 7.1: A model of a negative skewed curve



7.2.2 HCD Dimension

7.2.2.1 HCD Instructional Objectives

The teachers were found to be progressing in acquiring knowledge about HCD concepts and instructional objectives from the project's training, and were also found to be using instructional behavioural objectives from HCD levels in their lessons, which most of them stated well. This result was anticipated, since the sort of knowledge and skills included in this dimension are easily attainable. In addition, training efforts by

GMGEA mostly tackle this area due to its pivotal role in many subjects of education. Thus it either takes place as a training course dedicated to this theme or it forms an introductory section in courses on other themes. Moreover, the educational supervisors are well trained in this area and when visiting their teachers in schools, most of the time they repeatedly refer to this issue.

7.2.2.2 HCD Questions

As for HCD questions, the training course succeeded in improving the teachers' abilities in stating HCD questions to a fair level. After the training, they found utilising the HCD questions during their lessons, which reflects that the abilities acquired from training were sustainable. This is consolidated by the general interesting observation by the research assistants, that some teachers thought that the dimension of the HCD questions was the main aim of the project rather than the IAT skills – though they are both important to the project vision. This caused those teachers to focus on the HCD questions and give them additional focus. This interesting relationship between the two dimensions could be looked at in terms of what Tashkandi (1981) observed in his study, that the use of a practice-oriented instruction programme in question classification and questioning techniques significantly and positively affects the questioning behaviour of the teachers and increases the number of HCD questions used in the classroom. As presented in Chapter 1, it has been reported by a number of previous studies (Ayedh, 1993; Al-Darweesh, 1999; and Abdulmuttaleb, 1984) that in the Saudi context, little interest is given to HCD levels during instruction, in TA, and in textbooks, compared to the dominant interest in LCD. Consequently, the above-mentioned success of the researched teachers in the acquisition of HCD question construction skills is an important addition to the project's training, in relation to the current situation.

7.2.3 IAT Dimension

Using IAT requires initial success in running the CAIAT software package, thus the effectiveness in this respect has been examined. As answered by research question Q6 and by utilising various data collection methods, the observation reports showed that the teachers became capable of running the CAIAT easily and could utilise its main functions. The highest percentage was for running the software, which indicates that the electronic literacy barrier is not a concern. The four percentages next to the highest are each close to around 93.5%, which is not surprising since they all tackle how to enter

the data and print it out, which are two fundamental tasks for using the software. The two lowest percentages are for reviewing entered marks and editing errors in the data entry. Editing previous data tends to be a task that is more difficult than entering the data for the first time, because the training course begins by focusing on entering data and thus this, rather than editing the entered data[41], takes up most of the trainees' interest.

The ease of the teachers' use of the software is targeted by the design of its menus, screens, functions, and reports. This effort is because the CAIAT software is intended to be understood by the majority of regular users of personal computers. Fortunately, the good result found by the research confirms the feasibility of that effort. This finding relates to the similar intervention by Costagliola et al. (2007), who reported that the tutors, possessing no technical knowledge, were readily able to learn the new system and deal with it comprehensively. In addition, this is congruent with what Bermundo & Bermundo (2007) found in their study, that usability of their TCIAS system had a significance mean of 4.56 on a scale of 5.

As for IAT concepts and skills, although it is found by the pre-test that the teachers do not have a high level of previous knowledge about this area, the project, according to the workshop observations, post-test results and the CSs, has accomplished a good level in providing the participating teachers with this skill. The participants have achieved the required abilities in discovering weak or good items and to identify the underpinning reason, as well as to suggest how to overcome causing shortcomings. The lowest value of Table 6.13, which is item 12, indicates the more difficult task they dealt with, which is 'to compare the two IAT coefficients to each other to gain one judgement.' I believe that this is normal, since this type of skill is the most difficult among the rest of the IAT skills. Actually, it needs a considerable amount of practice and experience with different situations and pairs of numbers in order for the teacher to learn how to deal with these figures appropriately. Furthermore, the reported gradation in the difficulty of tasks is logical when we look at their function. I stress that these should be taken into account for future similar work, either in planning for training, software development or in follow-ups with teachers during the application.

As an overall conclusion, it could be stated that the CAIAT software package is similar to any common software package; therefore, most female science teachers, through short-term training, can easily learn CAIAT and understand its functions and menus. Another aspect is that IAT skills seem to be readily understood and applied by

the teachers. This is in line with previous research findings where Wang et al. (2004), Bermundo & Bermundo (2007), and Fan et al. (2011) have found that the use of computer-aided IAT was readily learnt by teachers and was perceived to function well in helping them to improve their abilities for better assessment practices. Furthermore, this result is in line with what the WATA system has achieved, according to the research by Wang et al. (2004), which reported that three items (see Table 1.3) were achieved positively: "to understand the strengths and weaknesses of test items," "to understand the strengths and weaknesses of choices for a multiple-choice item," and "to understand the strengths and weaknesses of a test."

I also believe that this could be interpreted in terms of Coniam's (2009) call to pay attention to empowering teachers to perform better in assessment. The CAIAT software fulfils the empowerment of teachers through facilitating and expediting their obtaining of difficult-to-calculate information needed for making decisions about their abilities. This is also congruent with what Henderson (1992) indicated that "Reflective teachers willingly embrace their decision-making responsibilities." He praised the fact that although teachers do not solve all of the problems they never stop trying. Decisions made by teachers about their practice represent an important pillar of AR methodology. In utilising the CAIAT for analysing teacher tests, the teacher, in fact, is following a kind of AR. The CSs' narrative presentation was intended to present a sample of the teachers' utilisation of this semi-AR practice during their self-learning of the CAIAT and their implementation of the underlying IAT.

7.2.4 Synopsis

- The teachers' previous background in HCD concepts and questions, and IAT skills, is limited.
- The limited levels of the teachers' abilities reflect that the pre-service training which is provided by colleges of education is very likely of a low level.
- The limited levels of the teachers' abilities compared to the high number of the teachers' years of experience reflect that the INSET provided by the MoE is very likely of a low level.
- The project training was effective both overall and for each of its content objectives (HCD and IAT).

- The teachers improved in their knowledge and skills about HCD concepts and instructional objectives, HCD question construction (during instruction and in tests), and IAT interpretation.
- The teachers easily became capable users of the CAIAT software and showed successful interpretations of the IAT parameters.
- In utilising IAT skills, discovering weak and good items is found to be somewhat easier than knowing what should be done after discovering those items; and comparing the two coefficients (difficulty and discrimination) to each other to gain one judgement is much more difficult.

7.3 Adoption Dimension

7.3.1 Adopting HCD

7.3.1.1 HCD Instructional Objectives

It has been shown, through item 3 of Table 6.17, that the teachers reported their interest in the area of HCD instructional objectives in the project. However, some further different answers for the sub-questions following item 3 need to be discussed in order to clarify how these answers were interpreted. As Table 7.2 shows, the reasons that were introduced, for those who answered 'No' to this question, are classified into two main categories: intrinsic reasons that come from the teachers' attitudes, understanding and abilities and hence reflect the functionality of the project as a change attempt (because it targets creating change within these characteristics), or external reasons that come from external sources and hence have no causal relationship with the project (because the project has no control over external reasons). These intrinsic and external reasons are listed in descending order for each in Table 7.2. The high percentages of the external reasons compared to those of the intrinsic reasons indicate that the group of sample teachers whose use of HCD instructional objectives did not increase were mostly willing to use HCD instructional objectives but that there were some external reasons that needed to be resolved so that they could do better in this respect. Among the reasons given were lack of time[42], the need for long-term training and perceived additional workload.

Table 7.2: Sorted reasons for not adopting HCD instructional objectives

Category	Reason	P
Intrinsic Reasons	e. I feel afraid that I might make some scientific errors if I tackled HCD, thus I tend to be limited to the lower cognitive level.	18.2%
	f. I think that teaching science/physics should not go as far as the level of higher cognitive demand.	0%
	g. I did not understand how I could apply HCD concept implications in the real world.	0%
	a. Not interested in this issue.	0%
	Sub total	18.20%
	Percentage of this section (with respect to the grand total)	25%
External Reasons	i. Other reasons: Because I have many subjects to teach.	36.4%
	h. Although I understand what I have learnt in this course, I think that I need some more time until I understand it thoroughly and am able to apply it.	27.3%
	c. This adds to my workload.	18.2%
	b. Do not have much time.	9.1%
	d. The school does not appreciate such improvement in my ability.	0%
	Sub total	54.60%
	Grand Total	72%
	Percentage of this section (with respect to the grand total)	75%

7.3.1.2 HCD Questions

Most of the sample teachers believe that the training provided by the project resulted in the number of their HCD questions being increased, either during instruction or in tests. As in Table 7.2, intrinsic and external reasons for item 4 of Table 6.17 are listed in descending order for each in Table 7.3, with very low percentages for the intrinsic reasons. This indicates that the group of the sample teachers whose use of HCD questions did not increase are very probably willing to do so but there are some external reasons that need to be resolved so that they can show a better response in this respect. Among the reasons given are a lack of time, a perceived additional workload, the need for long-term training and the appreciation from the school administration. The last of these is in line with what Betsworth (2003) found, that the head-teacher's support was pivotal for securing a sustainable PD (see Chapter 4).

Table 7.3: Sorted reasons for not adopting HCD questions

Category	Reason	P
Intrinsic Reasons	e. I feel afraid that I might make some scientific errors if I tackled HCD, thus I tend to be limited to the lower cognitive level.	14.3%
	a. Not interested in this issue.	0%
	f. I think that teaching science/physics should not go as far as the level of higher cognitive demand.	0%
	g. I did not understand how I could apply HCD concept implications in the real world.	0%
	Sub total	14.30%
	Percentage of this section (with respect to the grand total)	19%
External Reasons	i. Other reasons: Because I have many subjects to teach.	28.6%
	b. Do not have much time.	28.6%
	h. Although I understand what I have learnt in this course, I think that I need some more time until I understand it thoroughly and am able to apply it.	18.6%
	d. The school does not appreciate such improvement in my ability.	14.3%
	c. This adds to my work load.	0%
	Sub total	61.50%
	Grand Total	75.80%
	Percentage of this section (with respect to the grand total)	81%

7.3.2 Adopting the CAIAT Software and IAT Practice

As shown in Chapter 6, the teachers' initial interest in the CAIAT software was high. Also, their long-term PD is elicited by the CAIAT software. The following discussion aimed at elaborating about some difficulties that were mentioned in Table 6.18. Also, some negative responses mentioned by the sub-items of item 5 at Table 6.17 require further exploration.

Within the category of difficulties in Table 6.18, item 10 indicated the installation process, which was a major difficulty that affected the application intensively and consumed much of the technical support team's time[43]. In almost all of the cases at the pilot stage this happened before the new installation procedure was initiated which solved the problem substantially with the main sample's application. Nevertheless, item 6 of Table 6.4 reveals that installing a new software package was a new skill for most of the participants, albeit most of them knew how to use computers according to item 5 of Table 6.4. This could also aid in understanding the need for intensive technical support for the installation purposes[44]. Items 11, 12, 14 and 15 highlight the issue of the teachers' workload and time consumption if the number of pupils was high. This concern is a focal point in this study, as the other instruments' findings have revealed; for example, items b and f of Q5 in Table 6.17, which will be

explained shortly. Item 13 highlights the teachers' indication of the necessity of providing enough PCs in school so that this project can be applied. This recommendation is valid since the teachers were either using their own laptops or the school PC which is mostly dedicated for the school administrative tasks such as printing official letters and sending emails. They also indicated, with low percentages as in items 20-22, some other concerns that come from a normal pessimistic feeling towards the process of change, and these need no further comment.

As for item 5 in Table 6.18, intrinsic and external reasons for this item are listed in descending order in Table 7.4. The majority of the teachers associated their low interaction with CAIAT with external reasons, which indicates that they are mostly willing to adopt it but there are external reasons that need to be resolved first. These external reasons/factors do not negatively impact the adoption principle, because they are independent of the project's components, and are mainly: lack of time, perceived additional workload, and the need for long-term training and application. Furthermore, there are low percentages for the intrinsic reasons compared to the high percentages for the external reasons[45]. Intrinsic reasons could be neglected, due to their low average percentage (32%); nevertheless and because of the high level of importance of the present discussed dimension, I will comment on the most notable of these reasons in the paragraphs that follow, in order to scrutinise this aspect further and obtain a clearer picture.

Table 7.4, sub item i ("I do not agree that this is a way to improve test item construction") is at the top of the intrinsic reasons for not adopting CAIAT, which reveals that there is a negative attitude towards the CAIAT methodology. The second item ("I am very confident that my ability in writing test questions does not need any improvement") could also give a similar impression about their negative attitude. I think that this does not necessarily reflect a truly transparent answer. My appraisal utilises the notion of 'teacher acceptance' of new innovation. It is possible that at the beginning the teachers resist 'conceptually' because they consider that the new intervention could affect their work in a way that requires more work or responsibility. This issue is similar to a situation that happened in the movement of 'Language Across Curriculum', indicated in Chapter 3, where the teachers did not welcome the new trend conceptually because it added to their work volume. In addition, in the evaluative study by Al-Awwad et al. (2010a) for the Saudi national project of *Leading Schools*, the staff responses revealed a similar outcome[46]. The study reported an obstacle which was the

absence of incentives in the light of the perception that the new school model required an additional workload. For the present project, the highest ratio of the external reasons was that they do not have much time (item b) and next to that is the issue of their workload (item f). All of this strengthens the idea that they have the sense that CAIAT would add to their work volume, thus those who refused the project in the related items are very likely to prefer to justify their negative response by some conceptual reason such as the two top ones of internal reasons (items i and c). I illustrated earlier that using the expectancy theory as a motivational approach entails that people's involvement in change is a result of their positive expectations of personal benefits from that attempt at change. The opposite; in other words, negative expectations, will result in the refusal of the new change concept, as happened in this case.

Reflecting on the DoI theory model leads me to look at the percentage of almost 80% who tried out the new innovation and succeeded by no less than 60% (see Table 6.17), in a way indicating that the DoI theory's sub-rout of ‘continued rejection’ (Figure 3.7) is logically abandoned. The next logical step is to look at the indicators of the other alternative which is the users' ‘continual adoption’ as the model reveals. These are embodied in the users' valuable recommendations and comments as reported by the open-ended instrument and CSs. These have shown that the users are ready to adopt the innovation when their concerns are resolved. Therefore, their recommendations and comments could be considered as early indicators of a 'continual adoption' in the sense of the confirmed absence of 'continual rejection'.

Another side in this discussion is the *practical acceptance*. I indicate the “rational-empirical” strategy, which is one of the three strategies suggested by Dalin (1978) for bringing about change assuming that people are intelligent, rational and willing to adopt change. This is embodied in the reported aspects of the sample teachers' great enthusiasm and interest, and therefore indicates the high level of the rational-empirical trend. I will highlight some situations and events throughout the training and the contingent stage that reflect the extent to which the teachers appreciated the new intervention in practice and were willing to undertake it. Some examples are: installing the software on their laptops rather than on the school's PC, asking for extra time to use the school's PC in order to use the software, asking for help from their husbands or repairing their PC at a workshop when the installation failed, because they thought it was a problem with the machine. The research assistants’ team leader reported that many teachers had overcome obstacles that appeared to them and

sometimes paid money for this purpose; either for repairing their own computers or for asking an expert technician in a computer shop to install the software for them. However, I have to point out that although it was announced that technical assistance was available to them, they preferred to pay for quick service in order to have more time and gain more experience before the next session or meeting of the project[47]. One research assistant reported to me that:

> Teachers are enthusiast about using the software and getting their test results very quickly. They called me many times, asking about how to deal with some menus and functions and some told me that their colleagues were asking to be "taught" how to use the software and to learn what its function is. Furthermore, one teacher had already copied the software for some of the school teachers and trained a number of them. One school principal talked to me during my visit to the school, asking for this software to be given to all teachers and training to be provided to all, since schools suffer from the lack of good quality test construction.

All of these observations, along with the success that the effectiveness dimension revealed, outline the teachers' 'practical acceptance' of the new approach of CAIAT/IAT practice. When this 'practical acceptance' and the confirmed justifications of external reasons contrasted with the teachers' refusal position about the same concept (as item i or c of intrinsic reasons imply), the resulting paradox is logically interpreted by the perception that they could have self-defended their refusal position by attacking the concept of the project as stated by these intrinsic reasons, which undermines their significance. This triangulates the previous result of the project's success in the adoption dimension, as reported in Chapter 6.

Most importantly, the preceding discussion highlights the difficulty of accurately measuring adoption by a research setting that is limited in time and resources, such as the present one. This leads me to stress the need to evaluate adoption by a future better research approach and setting.

Table 7.4: Sorted reasons for not adopting CAIAT

Category	Reason	P
Intrinsic Reasons	i. I do not agree that this is a way to improve test item construction.	25.8%
	c. I am very confident that my ability in writing test questions does not need any improvement.	16.1%
	g. I did not feel confident that I would be able to apply what I learnt in this course.	9.7%
	a. Not interested in this issue.	6.5%
	e. I did not understand what the project was all about.	6.5%
	Sub total	64.60%
	Percentage of this section (with respect to the grand total)	32%
External Reasons	b. Do not have much time.	51.6%
	f. This adds to my workload.	38.7%
	h. I need some time until I understand thoroughly what I have learnt.	29%
	d. The school does not appreciate such improvement in my ability.	9.7%
	j. Other reasons: 1) Being new to using computers. 2) Because I have many subjects to teach.	6.5%
	Sub total	135.5%
	Grand Total	200.1%
	Percentage of this section (with respect to the grand total)	68%

7.3.3 Discussion of Adoption

The good level of adopting the two content objectives of this project reveals that the teachers interacted sufficiently actively with the project components. Because this research's prime aim is to evaluate utilising the CAIAT computer software in stimulating teachers' PD, positive findings in adoption dimension are direct reflection of achieving this aim. Therefore, it is important to shed light on this area in terms of what the theoretical framework reveals about the adoption practices and opinions of the participants. In addition, I will explain below how this project met its proposed theoretical model for change, the *personal power* model, illustrated in Chapter 3.

By using the notion of the "expert's paradigm," Ericsson (1996) differentiates between an expert and a novice-professional respectively: he sees that the expert relates to "forward-reasoning" for examining assumptions, whereas the novice relates to "backward-reasoning" (Gujski and Ben-Peretz, 2005). This is also stressed by other research works (Albanese and Mitchell, 1993; O'Donnell, 2004; and Shaukat, Arain, Alam and Shahid 2007). Forward-reasoning embodies induction, in which one works on the available data in order to find out what is unknown; whereas backward-reasoning embodies deduction, in which one starts from a hypothesis which is formulated for the purpose of looking for data that supports it (or maybe supports the opposite) (Crespo,

Torres, and Recio 2004 and O'Donnell, 2004). The teachers in the present research have used the "expert's paradigm," or the "forward-reasoning," in their analyses and interpretations of their findings, since they start from data and end with resulting reasons that explain this data, and then plan for the future responses accordingly. Using the experts' methodology in scrutinising their work is likely to provide teachers with the two emotional dimensions of being "comfortable" and "confident." As presented in Chapter 3, these were suggested by Eraut (2004a: 114) as additional dimensions to House's (1979) three dimensions that characterises all change processes and work at all levels, namely: technological, political and cultural. This is confirmed by what is reported by McNergney and Carrier (1980: 151) in the case of introducing teachers to certain innovations whereby, when their comfort level increases over time, teachers become more concerned about understanding how the learner is affected by the innovation, which adds positively to the merit of this sort of exercise. Furthermore, Eraut (2004a: 114) suggests that being "comfortable" aids the maintaining of the professional spread of activities and relationships and being "confident" encourages the taking up of new challenges. I think that both aspects are necessary for adopting and disseminating change. Furthermore, the case studies have shown the way the teachers have followed the AR settings in order to learn the new skills. This also brings forward what Eraut (1989: 184) calls the "personal theory," which results from the interaction between theory and practice. This could be seen in the teachers' expressions, justifications and recommendations for the situations under their investigation. Those who did not do well in the adoption dimension justified this lack mainly by external reasons. Although the percentages of these teachers are minimal compared to those who reported a good adoption, some of the external reasons raised by this group deserve elaboration as will follow; in order to connect these reasons/factors to what is reported by the relevant literature or previous research.

As highlighted above, it is notable from Tables 7.2 to 7.4 that the teachers raised two factors that mainly affected their adoption of HCD questions and CAIAT: *time* and *workload*. The reason "Do not have much time" was the first among other external reasons. Eraut stresses frequently in most of his writings that "time for reflection or explicit learning is difficult to find" (Eraut, 2005a: 1); hence demands of a busy, crowded workplace are the main contributor to decreasing the opportunity of learning for CPD (Eraut, 2004a: 114). Pelgrum's (2001) study of 26 countries revealed a similar concern about *time* as one of the top 10 obstacles raised by the teachers in those

countries. Moreover, the reason “I need some time until I understand thoroughly what I have learnt” takes second place in item 3 of Table 6.17; and the third place for items 4 and 5. I think that this subscribes to what Hejran (2002) found in his study about Saudi teachers, which revealed that their need for training in assessment come in the first place. This is why they ask for more time for further and more learning. Moreover, as suggested by Eraut (2005a), PD learning is characterised not as following stages of stopping points but rather as being on inter-related learning trajectories that do not necessarily progress onwards and upwards; hence I think that the *time* factor here could appear to be a permanent requisite.

The reason “This adds to my workload” was in third and second place for items 2 and 5 respectively. Previous Saudi studies have indicated this obstacle where a high teaching workload is considered one of the main obstacles hindering the optimum fulfilment of educational aims, as shown in Ayedh's study (1993), and affecting the optimum achieving of science teaching's aims, as in the study by Al-Darweesh (1999). It is not strange to find that teachers stress workload as an obstacle for a new intervention which in one way or another represents an additional chore for them. In this respect, I have highlighted in Chapter 3 about educational change that subcultures represented by professions or factions have their own micro-political role which could work for or against change (Eraut, 2004a: 113), thus complaints about the workload could be one way to fulfil the trend of the 'rejecting' micro-politics. Nevertheless, the Concerns Theory explains how, over time, these aspects disappear. It considers that, at the beginning, teachers pay close attention to personal aspects related to innovation. However, when these are resolved, teachers shift their attention and interest to the educational aims of the innovation (Vaughan, 2002). In addition, teachers might overestimate the volume of workload at the initial stages of application: Dougherty (1988) reported a low correlation of r=0.35 between objective workload and perceived workload (Winefield, 2003). This entails that only 35% of the objective workload meets the anticipations of the individuals, thus the most of that anticipated, almost 75%, is very likely not to take place in reality. If we consider these perspectives about growth of change viability, then the little signs of rejection that appeared in the teachers' justifications could be theoretically omitted. Nevertheless, in practice, these should be considered by avoiding their presence in following phases, so as to secure a much higher level of confidence in functionality. Moreover, this finding is not attributed to the nature of the CAIAT necessarily but most likely to the nature of the present project

being a pilot for the final proposed official implementation, thus the researched teachers might feel that they are doing additional work compared to their peers. In this respect, when this sort of practice becomes officially mandated and all teachers become required to deal with it, I think that this sort of feeling towards workload and school appreciation, though not significant at present, is very likely to disappear and then the teachers will shift their attention and interest to the educational aims of the innovation, as indicated by Concerns Theory.

I have considered that the *personal power* model, illustrated by Figure 3.2, is the conceptual personal change model which this project's design will follow in order to ensure the adoption of its new attempt at change. Next, I will refer to this model to review how this project fulfilled a fair level of the model's first two stages. For the first stage, *knowledge*, the model indicates that its two factors of 'fundamental knowledge' and 'motivation to learn' lead to the achievement of skills. As for the first factor, the comprehensive training with the printed material rich in the project's concepts and skills made the needed fundamental knowledge sufficiently available to trigger the teachers' enthusiasm to learn further. The second factor, 'motivators for change' results from the fact that this project acknowledged the teachers' contribution. This is achieved by the certificate given to the teacher at the end of the project's fieldwork period and also by making her school principal aware of the project aims and tasks, which helps the principal's appreciation of the extra effort that the teacher had made. Because the two factors of the *knowledge* stage of the model had been made available during the project's application, and especially as they are within the scope of the project, the research findings of the effectiveness dimension have shown clearly the good levels of achievement of the related skills, which confirms what the model suggests.

The second stage of the model, *attitude*, includes two factors, 'personal change' and 'adoption environment', both of which were considered by this project. The 'personal change' factors, as perceived by personal change models or theories, explained earlier in Chapter 3, were available. For example, the Exchange Theory stresses mutual trust and power of knowledge which are embodied in the project's training and the trust placed in the teachers to work individually in the contingent stage and provide their inputs. Also, the Expectancy Theory, especially Fullan's view of the feeling of accomplishment that derives from the excitement of being involved in a changing process, is obviously present within the attractive method of self-evaluation through the

scientific/normative way that the CAIAT provides. Next to the 'personal change' factor, the model requires the factor of 'adoption environment'. In this respect, a real effort was made in the project's design to make the teachers' participation easy and attainable. For example, the training has been done in work hours instead of in the evening, and counted as part of the teachers' PD official practices; the school PC has been made available to the teachers to utilise CAIAT with full technical support from the project's dedicated team members; and the educational supervisor has been part of this practice for encouraging the teachers to work on the CAIAT and pay attention to HCD during instruction. Furthermore, the project design has considered the different perspectives of PD such as what Vrasidas and Glass (2004) denoted in terms of teachers' learning being similar to their pupils' learning; what Eraut (2005a) refers to as "to start small but long," therefore, the content is confined to the focal theme of the project; what Argyris & Schon (1974) introduced in terms of double-loop learning, which is represented by the teacher's iteration in dealing with her IAT analysis results; and what is stressed by various writers about paying close attention to practice, which enriches reflection significantly. Hence, the level of 'personal adoption' that took place was a result of making these available along with the other predecessor factors mentioned earlier. This shows that the project has followed the personal power model to a good level.

The third stage, *individual behaviour*, I perceive, should be the business of the adopting organisation, as the MoE is recommended to undertake this project. This is because this stage depends mainly on the component of 'Organisational Reinforcement', which it is suggested the MoE should consider in its future wide scale application by providing the teachers with some kind of enforcement that encourages them to follow the new practice. I suggest that this enforcement is not necessarily a material-based one but even an encouraging evaluative scheme that acknowledges those who perform well in their TA. The result, as the model entails, is the final component of this stage, which is 'On-going Practice' and that is when the change aim is considered to be entirely accomplished by the majority of teachers and institutionalisation is then achieved.

7.3.4 Synopsis

- The teachers' use of HCD instructional objectives and the teachers' number of HCD questions, either during instruction or in tests, have both increased as a result of the project.

- The project was successful in triggering the teachers' interest in adopting the CAIAT software and IAT practice.
- Teachers' work time and workload represent the major external factors that are very likely to create a negative impact on the teachers' adoption of this attempt to change.
- No intrinsic factors pertaining to teachers' attitudes or abilities were found significant in negatively affecting their adoption.
- The first two main stages of the *personal power* model for change have been accomplished by this research project. The last stage is perceived to be accomplished through the recommended generalised wide-scale application.

7.4 Discussion of the Study Variables (for all dimensions)

The examination of statistical significance of any association of the teachers' researched variables with the two content objectives studied, HCD and IAT, revealed a number of results which will be elaborated shortly. Tables 7.5 and 7.6 summarise these findings for the two main dimensions of the study respectively: Effectiveness and Adoption. It is important to clarify that the associations between the researched variables and the research prime findings are found limited thus not impacting these findings. However, the following presentation is important to shed some light on the implications that these associations (or non-associations for some cases) reveal about the practices underpinning their existence (or non-existence).

Table 7.5: Summary of the researched variables' statistical significance for effectiveness dimension

Research Variables	Pre-Test			Post-Test		
	HCD		IAT	HCD		IAT
	Previous Know-ledge	Question Construc-tion		Previous Know-ledge	Question Construc-tion	
Level of Graduation	No	No	No	No	No	No
Educational Qualification	Yes	No	No	No	No	No
Years of Experience	No	No	No	No	No	No
INSET Courses on Test Construction	Yes	No	Yes	No	No	No
INSET Courses on IAT	Yes	No	No	No	No	No
Key Stage	Yes	No	No	No	Yes	Yes
Subject of Specialisation	Yes	Yes	No	No	No	No

Table 7.6: Summary of the researched variables' statistical significance for adoption dimension

Research Variables	HCD		IAT
	Use of HCD Objectives	Use of HCD Questions	Use of CAIAT Software
Level of Graduation	No	No	No
Educational Qualification	No	Yes	No
Years of Experience	No	No	No
INSET Courses on Test Construction	No	No	No
INSET Courses on IAT	No	No	No
Key Stage	No	No	No
Subject of Specialisation	No	No	No
Do you know how to use a computer?	No	No	No
Do you have a PC at home?	No	No	No
Can you use Excel software?	No	No	No
Can you use Access software?	No	No	No

7.4.1 Level of Graduation

Looking at the basic data of the teachers, their levels of graduation are on a normal distribution, whilst this variable has no association with any content objective of the project HCD or IAT. These two content areas are intended to be important parts of the educational subjects taught by colleges of education. Since they found having no association with the teachers' levels of graduation, this means that the colleges' educational subjects are very likely to make less contribution to the variance between the teachers' levels of graduation. As I have outlined before, the educational subjects provided by colleges of education do not seem to be achieving their intended aims and objectives as well as they should, and thus need to be reviewed in this regard.

7.4.2 Educational Qualification

Graduation from a college of education does not have any association except in relation to the teacher's background regarding HCD concepts. This seems logical, since the field of pre-service training iterates in providing teachers with this knowledge. It is the minimum anticipated result in this respect when we consider that 75% of the sample teachers are educationally qualified and that more than 75% of the sample teachers graduated with higher levels (very good to excellent) according to Table 6.2. Both factors should reflect the teachers' background in this area. In terms of adoption, this variable has shown an interesting association with adopting HCD question construction practice, where the teachers who graduated from colleges other than colleges of education have shown greater interest in adopting this sort of practice as a response to the project. This interesting relationship might be interpreted as meaning that these teachers lack this sort of background thus when this information was introduced to them they were more enthusiastic in applying the new knowledge**.**

7.4.3 Years of Experience

The number of years of experience has no association with any of the content objectives HCD or IAT. This sadly reveals that teachers do not make use of their experience over time; neither do they receive good INSET. I have highlighted in the first chapter that Al-Ahsa science education supervisors consider the lack of an authentic teacher performance evaluation as one of the reasons why teachers lack good question construction skills. Actually, most of the teachers receive high scores by this

evaluation without they spend great efforts that meet such score. This kind of 'work-culture' created their careless feeling towards INSET. I also explained in Chapter 4 the corresponding shortcoming in the *Form of Teacher Performance Evaluation*. On top of this is the fact that performance evaluation is not reflected in the teachers' annual salary increment and hence, as Friedman et al. (1980: 234) have pointed out, the “automatic” salary increment might have led to a decline in motivation to improve. This could explain why Saudi science teachers have not invested in their years of experience so far. On the other hand, this shows there is an urgent need to develop the educational supervision's practices so that educational supervisors disseminate skills and concepts that aid instruction and improve the teachers' outcomes. This could be achieved if some creative approaches for encouraging teachers were applied.

7.4.4 INSET Courses on Assessment

I have to highlight first the fact that the teachers attended two INSET courses on assessment: the first was on test construction and the second was on IAT skills. These were provided more than one year before this project by the training division of the GDGEA; therefore, they were part of the teachers' background and hence were considered among their researched characteristics. It was found that the two courses had an association with the teachers' background of HCD concepts in favour of those trained. This indicates that the two prior courses were targeting this skill and succeeded in accomplishing the training objectives in that area. However, the teachers' previous background regarding IAT was found to be associated with the training course on test construction only, in favour of those trained. Interestingly, the other training course, which was more related, because it was on IAT skills, had no significant impact. This contradiction led me to believe that of the two training courses, the course on test construction was successful and more comprehensive: at least to the extent of covering the IAT subject. The other course, though dedicated to IAT only, appeared not to function in terms of IAT. It did well in the area of HCD concepts only, which is usually an introductory course and represents the easiest and most attainable section of the content. This indicates that the IAT course was very likely less functional than the test construction course or rather weak. The head of the research assistants’ team, who is in charge of running such courses in her work, mentioned that the test construction course was provided first and that there were no IAT training courses given beforehand, thus it

included an IAT section. She also pointed out that it was successful and comprehensive compared to the IAT course, which confirmed my expectation.

7.4.5 Key Stage

The teachers who teach in secondary schools were better than those teaching in intermediate schools in terms of their HCD concepts background. This might be interpreted by what the diagrams 7.2 and 7.3 reveal where a high number of secondary school teachers have recently attended the INSET courses mentioned above, according to what the first column of each of the two diagrams shows. Moreover, the gradation of the number of participating teachers in these courses throughout the time period is characterised as decreasing for intermediate teachers and at the same time increasing for secondary teachers. This reveals that the planners of these courses were shifting their focus to secondary school teachers rather than intermediate school teachers and also it contributes to explaining the finding of post-test results for the HCD question construction and IAT skills, in which secondary school teachers were found to be much better at acquiring these skills. This indicates that secondary school teachers are more likely to succeed in achieving what the project implies, and this is in line with what is common among the educational supervisors and trainers, that these teachers tend to respond to PD programs much better than teachers of other key stages. This agrees with what Doran (1980: 64) highlighted, that writing HCD questions requires higher teacher ability, hence secondary school teachers are more likely to have better abilities in learning new ideas, dealing with issues that require thinking such as HCD question construction or making decisions such as IAT skills and adapting to change. Some Saudi researchers have also indicated this likelihood (for instance, Abdussalam, 1990).

Diagram 7.2: Comparison between intermediate and secondary school teachers' prior training in HCD concepts

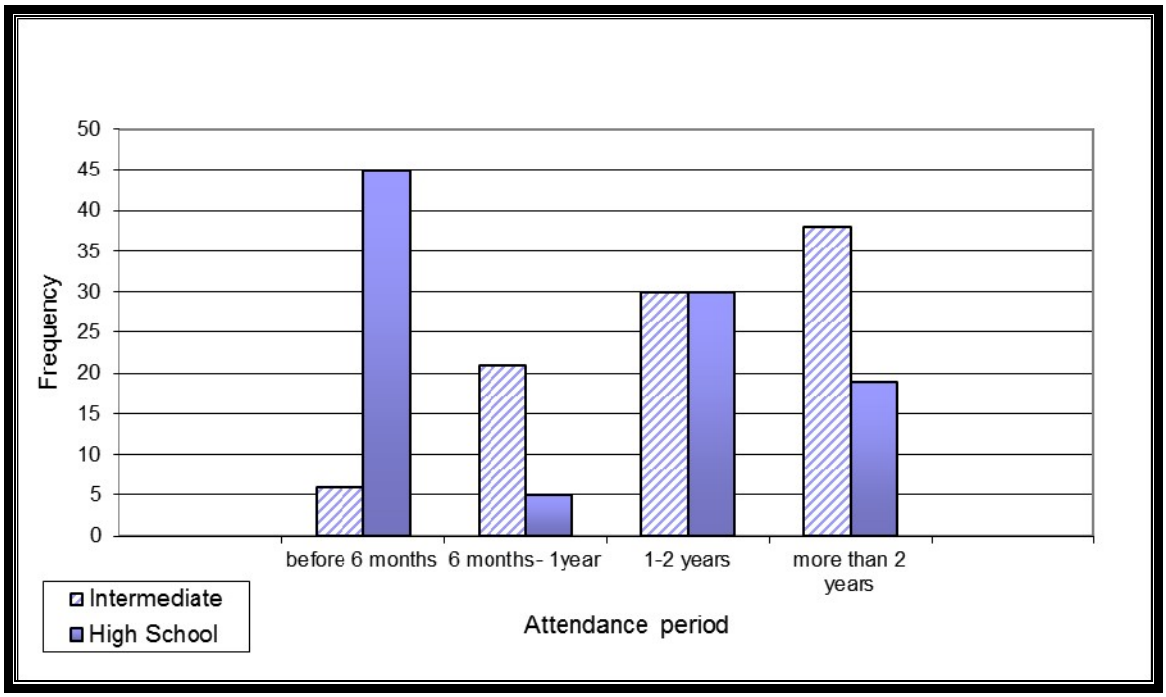
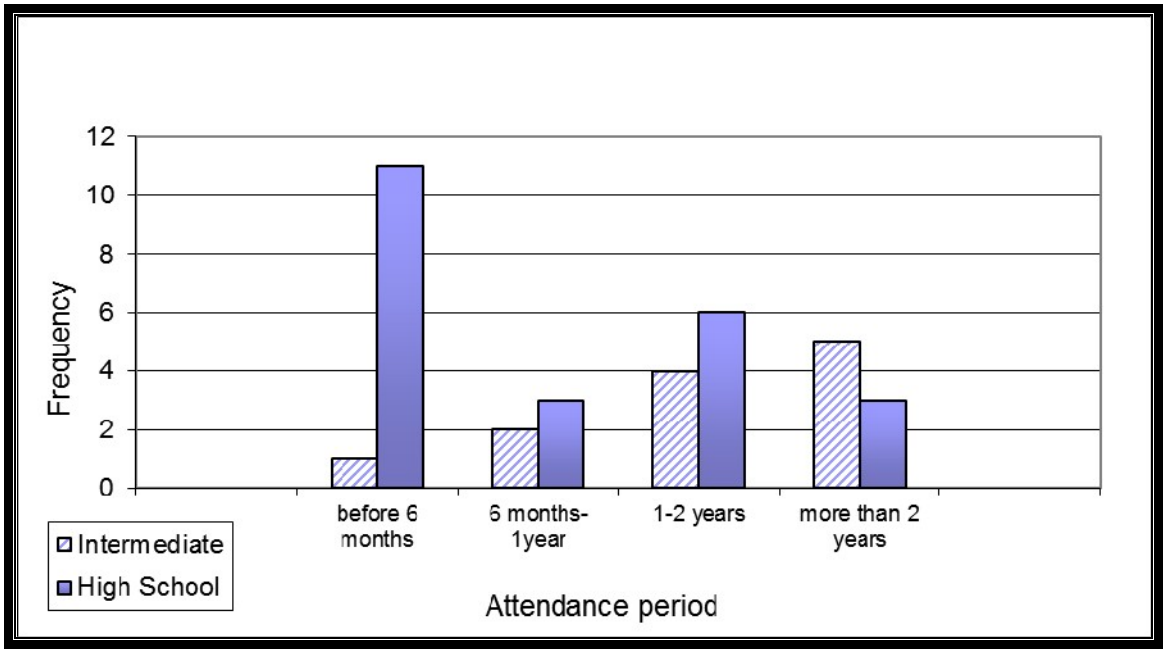


Diagram 7.3: Comparison between intermediate and secondary school teachers' prior training in IAT skills



7.4.6 Specialisation Subject

The association of a specialisation subject is found with the pre-test data for the HCD dimension only. Further analysis by Scheffe statistical treatment has shown that Biology and Physics are responsible for the associations with a background in the HCD concepts, and Chemistry and Biology for the association with the HCD test construction

skill. I do not have a comment here; because since these are related to background knowledge and skills only (pre-test) and I do not have previous studies that show a relationship between these specialisations and the skills under examination, then no solid interpretation could be addressed. However, this is of little importance to the discussion since post-test data have shown no association at all. Pre-test findings alone do not provide much valuable information for the aims of this discussion unless they show a relationship to post-test data. The function of post-test data is what we are seeking to explore in the first place; therefore, what counts more here is that because there is no association of HCD with post-test data I conclude that when training teachers in HCD, the specialisation subject does not need to be considered as a variable that could affect the projected outcomes.

7.4.7 Summary of the Research Variables

The different analyses of the study variables and their role in the study showed that, for the effectiveness dimension, most associations are within the HCD content objective for the previous knowledge of HCD while the other areas have limited associations. Most notably, the acquisition of all of the new skills (post-test section) is not associated with the researched variables except for the key stage. This overall independence of the acquisition of the project's prime two skills (HCD and IAT/CAIAT skills) from the teachers' researched characteristics opens the door to designing a large-scale training and dissemination strategies; or in other words: supports the generalisability of this study's findings. For the adoption dimension, none of the teachers' researched variables have any association with this area, which triangulates the likelihood of adoption success as revealed by this research's other findings and also supports the generalisability of this study's findings pertaining to adoption.

7.5 Conclusions

This research has found that the HCD/CAIAT package and project was effective in improving the Saudi female science teachers of the Al-Ahsa region in its two major content objectives, HCD and IAT. The first was represented by their achievement of knowledge pertaining to HCD concepts, and their skills and practices of writing HCD instructional objectives and HCD questions. The second was represented by their high level of ability in using CAIAT and their skills in utilising IAT successfully for assessing their test items.

Furthermore, this research has found that the HCD/CAIAT package was successful in encouraging the Saudi female science teachers of Al-Ahsa region to adopt its two content objectives HCD and IAT. In practice, this adoption is a result of the successful role of the CAIAT software in stimulating the teachers' PD for learning (on their own) how to improve their assessment skills for HCD levels. The good level of this project's effectiveness in achieving that the researched teachers acquired the researched skills with the triangulated indicators of the adoption dimension both work together in building confidence in the existence of the teachers' adoption for both HCD and IAT dimensions. To achieve the targeted adoption dimension, the project has followed a suggested 'personal power' model for change and was found feasible. The impact of the teachers' researched characteristics is found limited, and thus considered almost as not informing the overall conclusion.

All of the positive outcomes found by this research point to the likelihood of the anticipated impact that by following this project's proposal, the Saudi female science teachers' abilities/skills in HCD assessment will improve. This improvement should embody in their interest to pay greater attention to HCD instructional behavioural objectives and to write/ask good HCD questions for testing or during teaching practices which are core contributors to improving education and especially improving the quality of pupils' learning. However, the latter important issue was not tackled by the present research due to its limits; hence needs to be investigated through future long-term research. The project's success calls for generalising this research's findings to the population of the study, which is all of the female science teachers of intermediate and secondary schools in KSA with the importance of paying attention to the recommendations mentioned below when this project is widely applied.

7.6 Recommendations

As a result of the findings of this project, I recommend that the MoE in KSA adopt the HCD/CAIAT project to encourage all Saudi female science teachers to tackle HCD levels in their instruction and assessment. Furthermore, the level of adoption and enthusiasm for the project and its products seem to be an encouragement for the generalisation of this exercise to be extended to teachers of other subject areas and to a wider range of schools. I think that applying this project's package – the training and the CAIAT software – could add a great deal to the value of education in terms of the shift towards learning thinking and awaking creativity.

When the decision is made for this project to be adopted, it should be supported by a good preparation through training the trainers, improving the software, and printing manuals that aid a proper application and clarify the project's vision to all. I recommend that a future study should be applied to target hands-on details showing the extent to which this project could result in increasing the pupils' learning. This should be undertaken by MoE because on a large-scale and official study, teachers and principals will have a better attitude in participation. I recommend that a longitudinal study of no fewer than three years of data collection should follow up the project impact on teachers' adoption and pupils' learning. The latter could be measured by analyses of pupils' quality of answers and questions, pupils' comments during instruction, pupils' portfolios, case studies for some pupils, and pre/post-tests that are designed purposely to measure pupils' levels of various thinking styles/levels and progress of their academic achievement. Furthermore, designing such research using a quasi-experimental methodology where a control and experimental groups are studied over the years of the project is an optimal recommended approach.

Finally, the rest of my recommendations relate to the number of indications and comments that the participants and the fieldwork team have revealed for the sake of proper future application. Also, some recommendations are inferred from what some findings have shown about the actual situation of the researched teachers.

1. The workload of teachers should be looked at, so as to allow adequate time for PD practices other than instruction; among these: in-service training, PD workshops, teachers' meetings, IAT sessions and the like.
2. Application of the HCD/CAIAT package should consider the time-scale in the sense of giving adequate time for the training and follow-up by educational supervisors.
3. School principals and educational supervisors should understand the project's vision and aims, so that they cooperate and acknowledge the teachers efforts pertaining to this project.
4. Educational authorities that evaluate teachers' performance should acknowledge this sort of on-going PD by designing some form of follow-up and evaluative procedures that aid in identifying teachers' levels of interaction with this new trend and also address incentives that they could provide to outstanding teachers.

5. Pre-service training in Saudi colleges of education need to be studied through thorough research which aims at investigating the quality of these programmes in the related areas.
6. The training authorities in GDGEA (the researched LEA) are advised to increase their training in the areas of HCD and IAT since their teachers' years of experience did not show any significant differences in terms of their background in any of the topics studied.
7. The GDGEA should also look at the value of their previous training courses since some (about IAT) have appeared not effective.
8. The educational supervision practices should be developed to disseminate the skills of HCD and IAT to teachers and steer these skills to contribute in improving instruction.
9. The generalisation of the project should acknowledge the need to technical support for the CAIAT application especially at the beginning. I suggest online help, internet forum, and/or a phone call service.
10. The CAIAT software should be packaged in a technically professional way that eliminates any possible errors or dysfunctions that might appear during its installation or running.

7.7 Final Word

I need to highlight the outstanding benefits that I have reaped from carrying out this research, including the academic writing expertise that I have acquired. This is especially so because my work has been undertaken at a privileged British institution such as Sussex University. I have learnt how to write in a rhetorical way that strengthens my text, when to distil and when to expand, how to present the opinions of others within the whole meaning, and how to shift from one concept to another, linking the two as a series of related thoughts. There are some mistakes and errors that I have made during the fieldwork and writing up. Hopefully, I have discovered these, accepted the fact that I made them, and acted upon them, and this has strengthened my confidence in being able to cope with such situations in my future academic life. Moreover, I have developed my expertise in dealing with libraries, document centres, electronic libraries, databases, and the APA system. I have discovered that fieldwork is not as soft as I imagined. The discussions with my research assistants enlightened me in more than one area. Reflections that I have received from the pilot stage enabled me to

steer my thoughts about how to implement the fieldwork in the upcoming stage. Furthermore, as I have elaborated in Chapter 5, many lessons were learnt from this very critical and rich stage.

The inmost lesson that I have learnt in researching is that related to addressing the problem and positioning one's stand in that problem. This is because a problem has a variety of faces that do not appear at the beginning but gradually one discovers those faces when moving from one step to another, which sometimes causes puzzlement. I believe now that researchers need to spend a longer time on a problem before they attempt to research it. This is not simply in order to identify the problem but mainly to determine the best approach for dealing with its underpinning conceptual essences and practical variables.

I am a father of a large family and an LEA general manager who takes care of more than 100 thousand girl pupils in more than 350 girl schools alongside my membership of many MoE committees at the ministry level. Doing my research work whilst being bombarded with such a load of social and job roles has posed a great challenge along my way. In doing the research, I am driven by my interest and research criteria to produce quality work in a timely manner. I needed extensions more than once, albeit I had never imagined that I would. At times, I needed to separate myself from home and/or work, by travelling nearby for two or three days, to think thoroughly and focus on some important areas of the thesis. Taking on this challenge has contributed to shaping my life in a very serious way. It has made me a person who works diligently and makes use of every minute of his life, which I am happy with; because life is meaningless without an invention-oriented vision. Happily, I am proud of myself for dealing successfully with such a roaring challenge.

Bibliography

Abdulmuttaleb, M. (1984). A survey of some science teaching's problems in intermediate schooling at Riyadh (Masters dissertation). Riyadh, KSA, King Saud University.

Abdulrahman, S. (1983). *Psychological Measurement* (in Arabic). (1st Ed). Kuwait, Al-Falah Library.

Abdussalam, N. H. (1990). Trend of Saudi science female teachers in Makkah elementary schools towards teaching science and its relationship with level of qualification (Masters dissertation). Makkah, KSA, Um Al-Qura University.

Abo Hatab, F. & Othman, S. (1976). *Psychological Evaluation* [in Arabic]. (3rd Ed.). Cairo, Egypt, Egyption Anglo Publishers.

Adas, M. A. (1996). *School and teaching thinking*. Amman, Jordan, Dar Al-Fikr Publishers.

Adey, P. (1995). CASE: The long view, *CASE Network News*, Issue 2, London, UK, King's College.

Adey, P., Shayer, M. & Yates, C. (2001). *Thinking science package* [Teachers Guide]. (3rd Ed.). UK, Nelson Thornes limited.

Adkins, D. C. (1974). *Test construction: developmental and interpretation of achievement tests*. USA, Bell & Howell Company.

Al-Agha, A. R. (2004). Analysis of geography textbook of year 6 in palestine according to Bloom's taxonomy, *Journal of Islamic University, 12*, 451-467. Retrieved from http://www.iugaza.edu.ps/ARA/research/articles/volume12%20-%20Issue%202%20Human19.pdf, As visited on 3rd Jan 2009 at 8:45 p.m.

Al-Aklubi, M. (2008). Efficacy of using co-operative learning strategy in teaching "Hadeeth" subject for improving first year secondary pupils' academic achievement and critical thinking. (Phd Thesis). KSA, Um Al-Qura University. Retrieved online from: http://libback.uqu.edu.sa/hipres/ABS/ind5578.pdf, As visited on 1st Oct 2011 at 4:32 p.m.

Al-Awadh, Y. (1007). Efficacy of using HCD instructional strategy for teaching science on pupils' academic achievement and thinking development for year 6 pupils. (Masters' Dissertation). KSA, University of King Khalid. Retrieved online from:

http://libback.uqu.edu.sa/hipres/ABS/ind9968.pdf, As visited on 1st Oct 2011 at 5:05 p.m.

Al-Awwad, K., Abdul-Tawwab, A., Abu-Ouf, F., Helmi, F., Saddeeq, A., Shamrany, S., and Zaidan, H. (2010a). *Evaluating the project of "Leading Schools."* KSA, A governmental document prepared by Dar Masarat for Studies and Development. 1st edition, Riyadh: For the Ministry of Education, General Administration for Educational Research.

Al-Awwad, K., Abu-Ouf, F., Helmi, F., Redh, M., Shamrany, S., and Zaidan, H. (2010b). *Evaluating the project of "Attracting Schools."* KSA, A governmental document prepared by Dar Masarat for Studies and Development. 1st edition, Riyadh: For the Ministry of Education, General Administration for Educational Research.

Al-Bakr, M. (1998). Analytical evaluative study for physics tests' questions of final year of secondary school in KSA (Masters dissertation). Riyadh, KSA, King Saud University.

Albanese MA, & Mitchell S. (1993). Problem-based learning: a review of literature on its outcomes and implementation issues, *Academic medicine*. *Journal of the Association of American Medical College, 68*(7), 545. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8447896?dopt=abstract>, As visited on 29th Sep 2008 at 12:30 a.m.

Al-Dakheel, F (1997) Efficacy of employee performance evaluation form for female secondary school teachers at Riyadh (Masters dissertation). Riyadh, KSA, King Saud University.

Al-Damigh, K. & Al-Shumaimry, Y. (2010). Evaluation of teaching English language in the sixth grade of Saudi Arabian primary schools. Riyadh, KSA. A governmental document: Prepared by King Abdullah Institute for Research and Consultancy Studies, King Saud University, for the Ministry of Education, General Administration for Educational Research.

Al-Darweesh, N. (1999). The extent to which science teaching's aims are achieved in elementary boys schooling from the perspective of Al-kharj province's science teachers (Masters dissertation). Riyadh, KSA, King Saud University.

Al-Garfi, A (2010). Teachers' and pupils' perceptions of and responses to cooperative learning methods within the Islamic culture courses in one secondary school in Saudi Arabia (Phd Thesis). UK, University of Southampton.

Al-Harthy I. (2002). Project guide for Saudi leading schools, 2nd Edition. Riyadh, KSA. A governmental document. Centre for Educational Development, Ministry of Education.

Al-Khlaif, M. (2000). *Guide for leading schools, Chapter 2: Educational leadership and charter framework*. KSA, A government document. Retrieved online from: http://www.qassimedu.gov.sa/edu/showthread.php?t=5194, As visited on 21st Nov 2011 at 3:22 p.m.

Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory*. USA, Monterey, Calif.: Brooks/Cole Pub. Co.

Al-Mani', A. (2005). Learning styles preferred by middle school students in Riyadh, Saudi Arabia. DIRASAT: *Educational Sciences, 32*(2). Jordan. Retrieved online from: http://journals.ju.edu.jo/DirasatEdu/article/view/1539, As visited on 19th Sep 2011 at 6:32 p.m.

Al-Mazyed, M. (1975). Science education in public [state] secondary schools in Saudi Arabia as perceived by science teachers and science students (Doctoral dissertation). Ann Arbor, Michigan, USA, University of Oregon, Xerox University Microfilms.

Al-Mousa, N., Al-Abduljabbar, A., Al-Batal, Z, Al-Sartawi, Z., and Al-Husain, A. (2008). National research for evaluating KSA experience in mainstreaming SEN pupils in public [state] schools. KSA, A governmental document. Riyadh: Prepared for the Ministry of Education, General Administration for Educational Research.

Al-Nassar, S. & Al-Sughayyer, A. (2002) Instructional Practices of Teachers According to Learning Theories. Journal of Reading and Knowledge, 18. In Arabic. Retrieved Online from: http://faculty.ksu.edu.sa/dralnassar/dralnassar2/Lists/List/AllItems.aspx, As visited on 21st Sep 2011 at 13:37 p.m.

Al-Owaisheq, N. & Al-Suwailem, W. (2000). *Guide for leading schools*, *Chapter III: Module for curriculum and instruction/learning practices*. KSA. A governmental document. Retrieved online from: http://www.qassimedu.gov.sa/edu/showthread.php?t=5194, As visited on 21st Nov 2011 at 3:22 p.m.

Al-Saif, S. (1981). Recommended guidelines for the science education program in the state secondary schools of Saudi Arabia (Doctoral dissertation). An Arbor, Michigan, USA, University of Wyoming, University Microfilms International.

Journal for Educational Documentation (2003). Decree for teaching English in the sixth grade of primary school. *Journal for Educational Documentation (Attawtheeq Attarbawy), 48*, 1424 (in Hejri date), 2003. KSA, Ministry of Education.

Anderson, L. W. (1999). Rethinking Bloom's taxonomy: implications for testing and assessment [ERIC Database]. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/10/90/90.pdf, As visited on 3rd Sep 2005, at 3:43 a.m.

Antelmo, D. Costagliola, G. Ferrucci, F. & Fuccella, V (2005) A web-based computer aided assessment tool supporting question quality improvement. IADIS International Conference on Internet 2005. Dipartimento di Matematica e Informatica, Università di Salerno, Via Ponte Don Melillo. Retrieved online from: http://www.iadis.net/dl/final_uploads/200507L010.pdf, As visited on 28th Aug 2011, at 14:40 p.m.

Anzaldua, R. M. (2002). Item banks: Where, why, and how. Paper presented at the 25th annual meeting of the Southwest Educational Research Association, Austin, TX. Retrieved online from: http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/0d/c8/e2.pdf, As visited on 29th Aug 2011 at 17:15 p.m.

APA [American Psychological Association] (2010). Official website of APA, Retrieved from http://www.apastyle.org/learn/tutorials/basics-tutorial.aspx, As visited on 2nd July 2010, at 7:20 p.m.

APU instructional material (2009). Instructional material for MA in educational technology, Azusa Pacific University, Retrieved from: http://www.geibtechforlearning.org/apu/edu-522/blooms-taxonomy.jpg, As visited on 12th Nov 2009, at 1:20 p.m.

Aqeel, A. H. (1999). *Philosophy of inquiry methodologies*, Library, Egypt, Al-Madbooly.

Argyris, C., & Schon, D.A. (1978), *Organizational learning: A theory of action perspective*. San Francisco, CA, USA, Jossey Bass.

Ary, D., Jacobs, L. & Razavieh, A. (2004). *Introduction to research in education* [in Arabic]. Al-Ein, UAE, University Book Publishers.

ASC [Assessment Systems Corporation] (2006). *User's manual for the ITEMAN™ conventional item analysis program, for the 32-bit Windows© Version 3.6*, (2nd Ed.). St. Paul, Minnesota 55114, USA, ASC. *ITEMAN* is a registered trademark of ASC and *Windows* is a registered trademark of Microsoft Corporation.

ASC [Assessment Systems Corporation] (2009). Official website of ASC, St. Paul, USA, Retrieved from http://www.assess.com/xcart/skin1/html/compprog.htm, As visited on 17th Mar 2009 at 7:39 a.m.

Aseery, M. (1993). Relationship between level of thinking and level of academic achievement in some subjects for year 10 pupils according to Piaget stages of cognitive development (Masters dissertation). Riyadh, KSA, King Saud University.

Assyed, F. (1978). *Statistical psychology and measurement of human mind* [in Arabic]. (3rd Ed.), Cairo, Egypt, Dar Al-Fekr Al-Arabi.

Athanassiou, N., McNett, J. M., & Harvey, C. (2003). Critical thinking in the management classroom. *Journal of Management Education, 27*(5), 533-55, Retrieved from: http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&eric_viewStyle=list&ERICExtSearch_SearchValue_0=bloom&ERICExtSearch_SearchType_0=kw&eric_displayNtriever=false&eric_displayStartCount=21&_pageLabel=RecordDetails&objectId=0900000b80001063, As visited on 3rd Sep 2005, at 2:20 a.m.

Aviles, C. B. (1999). Teaching and testing for critical thinking with bloom's taxonomy of educational objectives, ERIC database, Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/23/26/9c.pdf, As visited on 4th Sep 2005 at 21:43 p.m.

Ayedh, A. (1993). The extent to which physics teaching's aims are achieved in secondary schooling from the perspective of physics teachers and educational supervisors at Riyadh (Masters dissertation). Riyadh, KSA, King Saud University.

Badr, A. (1989). *Essences of inquiry and its methodologies* [in Arabic]. Dar Al Maaref, Egypt.

Badr, B. (2006). Methods of teaching mathematics in girls' schools in Makkah Al-Mukarramah and its compatibility to modern age. Mission of Education and Psychology Journal, 26. (Resalat Al-Tarbiyah Wa Elm Al-Nafs), Published by GESTEN Association. Retrieved online from: http://uqu.edu.sa/files2/tiny_mce/plugins/filemanager/files/4281111/files/5.pdf And from: http://www.gesten.org.sa/portal/index.php?option=com_p7oth&task=show&catid=57&showid=116&Itemid=57, As visited on 1st Oct 2011 at 3:47 p.m.

Baez, A. V. (1970). *Modern trends in teaching science: Global perspective* [in Arabic]. (Salah Al-Ahmad & Shahadah Al-Khory, Trans.). Syria, Damascus University Printing Press.

Becker, H. J. (2000). Findings from the teaching, learning, and computing survey: is Larry Cuban right? School Technology Leadership Conference. Retrieved online from: http://www.crito.uci.edu/tlc/findings/ccsso.pdf, As visited on 12th Sep 2011 at 22:24 p.m.

Berman, M. L. (1971). *Motivation and Learning*. USA, New Jersey, Educational Technology Publications.

Bermundo C. B. & Bermundo A. B. (2007). Test checker and item analyzer with statistics. 10th National Convention on Statistics (NCS), EDSA, October 1-2, 2007, Retrieved online from: http://www.nscb.gov.ph/ncs/10thncs/papers/contributed%20papers/cps-13/cps13-01.pdf, As visited on 16th Aug 2011 at 21:06 p.m.

Best, J. W. (1981). *Research in education* [In Arabic]. (Abdulaziz Al-Ghanem, 1988. Trans.) Kuwait, Kuwait Corporation for Scientific Development.

Betsworth, M. (2003). Developing a whole school approach to the adoption of technology (Masters Dissertation). London, UK, King's College.

Bickman, L. & Rog, D. J. (1997). *Handbook of applied social research methods*. London, SAGE. Retrieved from http://books.google.com.sa/books?id=2A8w6KHJGIIC&printsec=frontcover&hl=ar&source=gbs_navlinks_s&redir_esc=y#v=onepage&q=case%20study&f=false, As visited on 12th Dec 2009 at 5:33 p.m.

Bill, J. M. (1977). Effects of varying structure and method on the validity of self-report measures, *British Educational Research Journal, 3*(1), 19-21. Available from

http://dx.doi.org/10.1080/0141192770030107, As visited on 22nd March 2008 at 4:12 p.m.

Black, P., Harrison, C., Bethan Marshall, C. L. & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom, Phi Delta Kappan, Questia Media America, Retrieved from http://www.questia.com/googleScholar.qst;jsessionid=J6MJL5pjFFF6K0pkZQ8GnQNz1jQ6Vvsp3fdDRqlGkyXyz1rVMBL2!1083457915!-1477775285?docId=5007120502, As visited on 13th Mar 2009 at 8:28 p.m.

Bloom, B. S., Engelhart, M. D., Hill, W. H., Furst, E. J. & Krathwohl, D. R. (1985). *System for classification of educational goals* (Vol. 1). (1st ed.), (Khawaldah, Mohammed Mahmood, & Awdah, Sadeq Ibraheem, Trans.). Jeddah, KSA, Shorooq Publication house.

Bloom, B. S., Madaus, G. F., & Hastings, J. T. (1981). *Evaluation to improve learning*. USA, McGraw-Hill Book Company.

Boeree, C. G. (2006). Jean Piaget: biography, Retrieved from Boeree webpage http://www.ship.edu/~cgboeree/piaget.html, Part of the Shippensburg University website http://ship.edu/, As visited on 16th October 2006 at 2:25 p.m.

Bond, L. (1994). Reaching for new goals and standards: The role of testing in educational reform policy, excerpted from NCREL's policy talks, Audiotape2, Audio comment, 204k, As cited in NCREL's website, Retrieved from www.ncrel.org\sdrs\areas\issues\methods\assment\as700.htm, As visited on 23rd Nov 2002 at 4:50 p.m.

Bowers, C. A. (2000). Let them eat data: How computers affect education, Cultural diversity, and the prospects of ecological sustainability. USA, Athens, The University of Georgia Press. Retrieved online from: http://books.google.com/books?printsec=frontcover&vid=ISBN0820322296&vid=ISBN082032230&vid=LCCN00026718#v=onepage&q&f=false, As visited on 12th Sep 2011 at 23:43 p.m.

Brikeet, A. (2009). Efficacy of educational units depending on constructivist learning model for improving grammar skills and attitudes of first secondary year pupils. (Phd Thesis). KSA, University of Umm Al-Qura. Retrieved online from: http://libback.uqu.edu.sa/hipres/ABS/ind7476.pdf, As visited on 1st Oct 2011 at 4:39 p.m.

Brooks, J. G. (2002). Schooling for life: Reclaiming the essence of learning, Published online by Association for Supervision and Curriculum Development (ASCD), Retrieved from http://www.ascd.org/portal/site/ascd/template.chapter/menuitem.b71d101a2f7c208cdeb3ffdb62108a0c/?chapterMgmtId=f4c1177a55f9ff00VgnVCM1000003d01a8c0RCRD.

Brown, F. G. (1970). *Principles of educational and psychological testing*. USA, The Dryden Press Inc.

Brown, J. C. Frishkoff, G. & Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. In proceedings of HLT/EMNLP 2005. ACL Anthology, A digital archive of research papers in computational linguistics. Retrieved online from: http://aclweb.org/anthology/H/H05/H05-1103.pdf, As visited on 28th Aug 2011 at 16:21 p.m.

Brown, P. J. (2003). Analysis of an innovation from the perspective of change [MA dissertation]. UK, University of London.

Brownstein, E. M. (1997). Interaction between assessment and instruction in the science: A teacher's decision-making process (Doctoral dissertation). USA, The Ohio State University. Source: ProQuest search engine, Digital dissertation division, Retrieved from: http://www.lib.global.umi.com/dissertations/fullcit/9813228, As visited on 9th Nov 2004 at 12:26 p.m.

Butterfield, Sue (1995). *Educational objectives and national assessment*. Buckingham, Open University Press.

Bybee, R. W. & Loucks-Horsley, S. (2000). *Advancing technology education: The role of professional development*. The technology teacher, Retrieved from http://www.iteaconnect.org/TAA/LinkedFiles/Articles/TTTpdf/2000-01Volume60/bybeehorsley_oct00.pdf, As visited on 27th Aug 2008 at 8:12 p.m.

Cangelosi, J. S. (1982). *Measurement & evaluation: An inductive approach for teachers.* USA, WM. C. Brown Company Publishers Ltd.

CERIS –PAC [Research Training Project] (2004). Introduction to research methodology in immigration and settlement, Retrieved from: http://ceris.metropolis.net/pac/pac13.pdf, As visited on 30th Dec 2004 at 6:20 p.m. Also available from: http://www.keepandshare.com/doc/1629574/pac09-pdf-december-21-2009-9-29-pm-257k?dn=y, As visited on 6th Dec 2011 at 4:18

p.m. This is a workshop material by the same author titled "Introduction to Action and Participatory Research"

Child, D. (1983). *Psycology and the teacher* [In Arabic]. (Elsayed, Abdul halim Mahmoud, Darwish, Zein El Abedeen & El Dureiny, Hussein Abdel Aziz, Egypt, Al Ahram Est. Trans.) Eastbourne, England, Holt Saunders Limited.

Clare, J. (2005, Mar 21st). Pupils make more progress in 3Rs 'without aid of computers'. The Telegraph Newspaper. Retrieved online from: http://www.telegraph.co.uk/news/uknews/1486108/Pupils-make-more-progress-in-3Rs-without-aid-of-computers.html, As visited on 12 Sep 2011 at 21:54 p.m.

Clift, J. C. & Imrie, B. W. (1981). *Assessing students, appraising teaching.* London, UK, Croom Helm Ltd.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education.* London, UK, Routledge Falmer.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education. 6th edition,* London, UK, Routledge Falmer. Retrieved online from: http://books.google.com.sa/books., As visited on 27th Dec 2011 at 4:35 p.m.

Cole R. & Williams D. (1973). Pupil responses to teacher questions: cognitive level, length and syntax. Published online by Association for Supervision and Curriculum Development (ASCD). Retrieved online from: http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_197311_cole.pdf, As visited on 31st Aug 2011 at 02:11 a.m.

Collins, H., Johansen, J. H. & Johnson, J. A. (1976). *Educational measurement and evaluation: A work text*. USA, Scott, Foresman and Company.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the effects of training in test development principles and practices on improving test quality. *System, 37*(2), 226-242,. Retrieved online from: http://www.sciencedirect.com/science/article/pii/S0346251X09000062, As visited on 1st Sep 2011 at 16:48 p.m.

Comley, A. D. (2001). A case study looking at assessment and reporting arrangements in a secondary school setting using a computerised record keeping system (MA dissertation). London, UK, King's College.

Coolidge, S. W. (1989) Using higher order cognitive questions in the primary classroom to improve comprehension. Dissertations/Theses - Practicum Papers, ERIC

database, Retrieved from http://www.eric.ed.gov, As visited on 3rd Sep 2005 at 2:06 a.m.

Costagliola, G., Ferrucci, F., Fuccella, V., Oliveto, R. (2007). eWorkbook: a Computer Aided Assessment System. *International Journal of Distance Education Technologies, 5(3)*, 24-41. Retrieved online from: http://weblab.dmi.unisa.it/weblab/images/stories/papers/jdet07.pdf, As visited on 10th Sep 2011 23:33 p.m.

Costagliola, G. & Fuccella, V. (2009). A Rule-Based E-Testing System for Test Quality Improvement. *International Journal of Distance Education Technologies 7*(2), 63-82, Retrieved online from: http://weblab.dmi.unisa.it/weblab/images/stories/papers/jdet09.pdf, As visited on 10th Sep 2011 at 22:50 p.m.

Cotton, K. (1988). Classroom questioning. U.S. Department of Education, A publication based on work sponsored wholly, or in part, by the Office of Educational Research and Improvement (OERI), under Contract Number 400-86-0006. Also, NorthWest Regional Educational Laboratory, Annenberg Learner, The Annenberg Foundation. Retrieved online from: http://www.learner.org/workshops/socialstudies/pdf/session6/6.ClassroomQuestioning.pdf, As visited on 31st Aug 2011 at 00:18 a.m.

Cotton, K. (2011). Questioning techniques. USA, Annenberg Foundation, Insights into Algebra 1: Teaching for learning, an eight-part video, print, and Web-based professional development workshop for in-service teachers. Retrieved online from: http://www.learner.org/workshops/algebra/about.html, As visited on 31st Aug 2011 at 01:45 a.m.

Crespo, K. E., Torres, J. E., & Recio, M. E. (2004). Reasoning process characteristics in the diagnostic skills of beginner, competent, and expert dentists. *Journal of Dental Education, 68*(12), 1235-1244. Retrieved from http://www.jdentaled.org/cgi/content/full/68/12/1235, As visited on 28th Sep 2008 at 5:38 p.m.

Cuban, L. (2001). Oversold and underused: computers in the classroom. USA, Harvard University Press. Retrieved online from: http://www.hull.ac.uk/php/edskas/Cuban%20article%20-%20oversold.pdf, As visited on 12th Sep 2011 at 22:43 p.m.

Custer, R. L. (1996). Qualitative research methodologies. *Journal of Industrial Teacher Education, 34*(2), 3-6. Retrieved online from: http://scholar.lib.vt.edu/ejournals/JITE/v34n2/editor.html, As visited on 21 Feb 2012 at 13:30 p.m.

Dalen, D. (1962). *Understanding Educational Research: An Introduction*. USA, New York, McGraw-Hill. Arabic Version, Translated by: Nawfal, M., al-Shaikh, S. and Gabriel, T. (1990). Egypt, Cairo, Egyption Anglo Publishers.

Daniel L. g. & King D. A. (1998). Knowledge and Use of Testing and Measurement Literacy of Elementary and Secondary Teachers. The Journal of Educational Research, 91(6), pp. 331-344. Taylor & Francis, Ltd. Retrieved online from: http://www.jstor.org/stable/27542177, As visited on 6th Aug 2011 at 7:36 a.m.

David, M. (2007). Website of Oregon Technology in Education Council (OTEC). A non-profit organisation, USA. The website is update 27th Feb 2007. Retrieved online from: http://otec.uoregon.edu/arguments_against.htm, As visited on 12th Sep 2011 at 21:13 p.m.

Davies, B. (2004). Exploring the tensions between how teachers want to teach and how they are required to teach in order to fulfil what is required of them by government, schools, pupils and parents (MA dissertation). London, UK, King's College.

Davis, T. N. (1975). *Objective tests in theory and practice*. Reduit, Mauritius Institute of Education.

Dean, J. (1991). *Professional development in school*. Buckingham, UK, Open University Press.

Dillon, A. & Morris, M. G. (1996). User acceptance of new information technology: theories and models. In Williams, Martha E., (Eds.) *Annual Review of Information Science and Technology, [chapter 31]*, pp. 3-32. Medford, N.J.: Information Today. Retrieved from http://www.ischool.utexas.edu/~adillon/BookChapters/User%20acceptance.htm, As visited on 22nd Aug 2008 at 10:16 p.m.

Doran, R. L. (1980). *Basic measurement and evaluation for education* [In Arabic]. (Sabarini, Mohammed, AL-Khalily, Khalil, & Malakawy, Fathi, Trans.). NW, Washington, DC, USA, Science Teachers Association 1742 Connecticut Avenue.

Driver, R. (1983). *The pupil as scientist?* UK, The Open University Press.

Ebel, R. L. (1972). *Essentials of educational measurement.* Englewood Cliffs, New Jersey, Printice-Hall, Inc.

EIA (2001). The Energy Information Administration website. A statistical agency of the U.S. Department of Energy. Retrieved from http://www.eia.doe.gov/emeu/cabs/saudi.html, As visited on 6th Nov 2001 at 10:30 a.m.

Eggen, T. J. H. M. & Straetmans, G. J. J. M. (2000) Computerized Adaptive Testing for Classifying Examinees into Three Categories. *Educational and Psychological Measurement, 60*(5), 713-34, Retrieved on line from: http://www.eric.ed.gov/, As visited on 29th Aug 2011 at 17:45 p.m.

Eraut, M. (1989). *Initial teacher training and the NYQ model, competency based education and training*. Burke, John W. Routledge. Retrieved from http://books.google.com, As visited on 3rd Nov 2006 at 6:55 p.m.

Eraut, M. (2004a). Editorial: Learning to change and/or changing to learn. *Learning in Health and Social Care, 3* (3), P. 111. Retrieved from http://www.blackwell-synergy.com/doi/full/10.1111/j.1473-6861.2004.00073.x?prevSearch=, As visited on 3rd Nov 2006 at 4:33 p.m.

Eraut, M. (2004b). Editorial: The practice of reflection. *Learning in Health and Social Care, 3*(2) P. 47. Retrieved from http://www.blackwell-synergy.com/doi/full/10.1111/j.1473-6861.2004.00066.x?prevSearch=, As visited on 3rd Nov 2006 at 4:05 p.m.

Eraut, M. (2005a). Editorial: Continuity of learning. *Learning in Health and Social Care, 4*(1), P. 1. Retrieved from http://www.blackwell-synergy.com/doi/full/10.1111/j.1473-6861.2005.00086.x?cookieSet=1, As visited on 6th Nov 2005 at 15:95 p.m.

Eraut, M. (2005b). Professional knowledge in medical practice. *La Profesion Medica: Los Retos del Milenio*, Israel, Website of Mofet Institute. Retreived from http//www.mofet.macam.ac.il/iun-archive/.../ProfessionalKnowledgeinPractice.pdf, As visited on 1st July 2010 at 5:30 p.m.

Fairon, C. (1999). "A Web-based System for Automatic Language Skill Assessment: EVALING". Proceedings of Computer Mediated Language Assessment and Evaluation in Natural Language Processing Workshop. Retrieved online from:

http://acl.ldc.upenn.edu/W/W99/W99-0410.pdf, As visited on 28th Aug 2011 at 15:50 p.m.

Fan, Y.C., Wang, T.H. & Wang, K.H. (2011) A Web-Based Model for Developing Assessment Literacy of Secondary In-Service Teachers, *Computers & Education Journal, 57*(2), 1727-1740, Sep 2011, Retrieved from http://www.eric.ed.gov, As visited online on 4th Aug 2011 at 23:50 p.m.

Ferguson, S. (2005). How computers make our kids stupid: There's growing evidence that too much cyber-time dumbs down our children. McCleans. Retrieved online from: http://www.macleans.ca/education/universities/article.jsp?content=20050606_106930_106930, As visited on 12th Sep 2011 at 23:54 p.m.

Ferreira, F. & Santos, J. N. et al. (2007). Information professionals in Brazil: core competencies and professional development. *Information Research, 12*(2), P. 299. Retrieved from http://informationr.net/ir/12-2/paper299.html, As visited on 22nd Aug 2008 at 10:25 p.m.

Fielding, N. & Schreier, M. (2001). Introduction: on the compatibility between qualitative and quantitative research methods, *Forum Qualitative Sozialforschung* / *Forum: Qualitative Social Research* [On-line Journal], *2*(1). Retrieved from http://www.qualitative-research.net/fqs-texte/1-01/1-01hrsg-e.htm, As visited on 29th Dec 2004 at 10:27 p.m.

Fink, A. G. (2002). *How to sample in surveys: The survey kit 7*, USA, SAGE-USA. Retrieved from http://books.google.com, As visited on 12th Dec 2009 at 3:18 p.m.

Fitch, M. E., & Kopp, O. W. (1990). *Staff development for the practitioner.* USA, Charles C Thomas Publisher.

Freeman, R. & Lewis, R. (1998). *Planning and implementing assessment.* London, Kogan Page Limited.

Friedman, M. I., Brinlee, P. S., Hayes P. B. D. (1980). *Improving teacher education: Resources and recommendations.* New York, Longman.

Fullan, M. (2001). *The new meaning of educational change.* London, UK, Routeldge Falmer.

Gage, N. L., & Berliner, D. C. (1984). *Educational psychology*, (3rd Ed.). USA, Houghton Mifflin Company.

GDBEA [GDBEA: General Directorate of Boys Education at Al-Ahsa] (2005). An official letter from the director of GDBEA. [As a response to my query letter about the latest statistic of secondary schools Physics teachers in Al-Ahsa Province boys schools].

Ghareeb, Ramzeyyah (1985). *Psychological and educational evaluation and measurement* [In Arabic]. Cairo, Egypt, Egyption Anglo Publishers.

Gipps, C. V. (1994). B*eyond testing: Towards a theory of educational assessment.* London, UK, Routledge Falmer.

Glaser, B. & Strauss, A. (1967). *The Discovery Of Grounded Theory: Strategies For Qualitative Research*, USA, Transaction Publishers. Retrieved online from: http://books.google.com, As visited on 13 Nov 2011 at 6:44 a.m.

Gobo G. (2007). Sampling, Representativeness and Generalizability. Website of University of Milan. Retrieved online from: http://ggobors.ariel.ctu.unimi.it/repository/ggobors/sage2004.pdf, As visited on 12 Nov 2011 at 01:12 p.m.

Gujski, J. & Ben-Peretz, M. (2005). Professional-development and professional-uncertainty of teachers, Paper presented at the British Educational Research Association annual conference, PP. 14-17, Pontypridd, UK, University of Glamorgan. Retrieved from http://www.leeds.ac.uk/educol/documents/153028.htm, As visited on 29th Aug 2008 at 11:42 p.m.

Gulf Law (2011). The website of Gulf Legal Services Ltd. 30 Kingston House South, Ennismore Gardens, London. UK. Retrieved from: http://gulf-law.com/saudi_map.html, As visited on 7th Feb 2012 at 11:26 p.m.

Guskey, T. R. (2000). *Evaluating professional development.* California. USA, Corwin Press Inc. Retrieved from http://books.google.com,M1, As visited on 15th Aug 2008 at 8:03 p.m.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. USA, SAGE.

Hambleton, R. K. (1984). Using Microcomputers to Develop Tests. *Educational Measurement: Issues and Practice, 3*(2) 10–14. Retrieved online from: http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1984.tb00742.x/abstract, As visited on 16th Aug 2011 at 17:38 a.m.

Hamdan, M. Z. (1998). *The Dialogue and classroom questions: Provoking thinking through education*. Damascus, Syria, Modern Education House.

HCEP: Higher Committee for Education Policy in KSA (2006). Pupil Assessment Document. Riyadh, KSA, HCEP.

Hejran, A. M. (2002). Descriptive study to determine training needs for teachers: An introduction for constructing proposed training programme from the perspective of educational leaders, professionals, and educational supervisors (in Arabic). *Journal of Um Al-Qura University for Educational, Social and Human Sciences, 14*(1). Makkah, KSA.

Henderson, J. G. (1992). *Reflective teaching becoming an inquiring educator*. USA, Macmillan Publishing Company.

Hersey, P. & Blanchard, K. (1988). *Management of Organizational Behaviour: Utilizing Human Resources*, USA, Prentice-Hall International, Inc.

Hills, J. R. (1981). *Measurement and evaluation in the classroom*, Columbus, Charles E. Merrill Publishing Company.

Hodges, M. & Lachs, J. (2000). *Thinking in the Ruins*. Nashville, Vanderbilt University Press.

Holbrook, J. (2003). Rethink Science Education. *Asia-Pacific Forum on Science Learning and Teaching, 4*(2). Foreword (Dec, 2003). Retrieved from http://www.ied.edu.hk/apfslt/v4_issue2/foreword/index.htm, As visited on 27 Oct 2006 at 4:10 p.m.

Howell, D. C. (2007). Treatment of missing data. Howell web page, Retrieved from http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html, Last update 3/23/2007, As visited on 20 Sep 2007 at 01:53 a.m.

Hummel, J., & Huitt, W. (1994). What You Measure is What You Get. GaASCD Newsletter: The Reporter, 10-11. Retrieved online from: http://www.edpsycinteractive.org/papers/wymiwyg.html, As visited on 3rd Sep 2011 at 10:50 p.m.

Jaworski, B. (1995). A Book Review of: Constructivism in Education. by Steffe, L. P. & Gale, J. Published 1995, UK, Oxford. Part of the European Mathematical

Information Service. Retrieved from: http://www.emis.de/journals/ZDM/zdm982r2.pdf, As visited on 20th Oct 2006 at 3:16 a.m.

Johnson, G. (1986). Criterion-referenced assessment. In: Lloyd-Jones, Robin & Bray, Elizabeth (1986). *Assessment from principles to action*. Hampshire and London, MacMillan Education Ltd.

Jones L. & Fletcher C. (2004) The impact of measurement conditions on the validity of self-assessment in a selection setting. *European Journal of Work and Organizational Psychology*, *13*(1), 101–111. Available from http://pdfserve.informaworld.com.ezproxy.sussex.ac.uk/794956_751307726_744085204.pdf, As visited on 22nd March 2008 at 3:30 p.m.

Kearns, S. P. (not dated). Scoring, item analysis, reliability, and validity [PowerPoint slides]. Charleston, SC, USA, Trident Technical College.

Keefe, J. W. & Walberg, H. J. (1992). *Teaching for thinking*, USA, National Association for Secondary School Principals. (Arabic version, 1995), Riyadh. KSA, Arab Bureau of Education for the Gulf States.

Kehoe, J. (1997). Basic item analysis for multiple-choice tests. ERIC Digest, Retrieved from http://www.ericdigests.org/1997-1/basic.html, As visited on 17 Jan 2004, at 22:00 p.m.

Kelle, U. (2001). Sociological explanations between micro and macro and the integration of qualitative and quantitative methods [43 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [online journal], 2*(1). Retrieved from http://www.qualitative-research.net/fqs-texte/1-01/1-01kelle-e.htm, As visited on 30th Dec 2004 at 11:45 a.m.

Kim, Seock-Ho (1999) A Computer Program for Classical Item Analysis, The University of Georgia, Published online on Jun 18, 1999, Retrieved from <http://www.arches.uga.edu/~shkim/epm99.htm>, As visited on Thu 11 July 2002, at 8:57 p.m.

Kristinsdóttir, S. B. (2001). Lev vygotsky. An article in Síðast uppfært web site, Kristinsdóttir is a Project Manager of the doctoral study and in the dep. of Continuing Education at the School of Education, University of Iceland. Dated 08.11.2001, Retrieved from http://starfsfolk.khi.is/solrunb/vygotsky.htm, As visited on 20th Oct 2006 at 9:36 p.m.

Koontz, H. & O'Donnell, C. (1976). *Management: A Systems and Contingency Analysis of Managerial Functions*, (Sixth Edition), USA, McGraw-Hill, Inc.

Kumar, David (1997). Computers and Assessment in Science Education. ERIC Digest. Retrieved online from: http://www.ericdigests.org/1997-1/science.html, As visited on 17th Jan 2004 at 22:00 p.m.

Kumar, R. (2005). *Research methodology: a step-by-step guide for beginners*. SAGE. Retrieved online from: http://books.google.com/books?id=x_kp__WmFzoC, As visited on 23rd Sep 2011 at 11:06 a.m.

Leat, D. (1993) Competence, Teaching, Thinking and Feeling. *Oxford Review of Education*, 19(4), 499-510. Available from http://www.jstor.org.ezproxy.sussex.ac.uk/stable/pdfplus/1050568.pdf, As visited on 17 Dec 2010 at 9:19 p.m.

Ledda, M. (2005). A French lesson: Three new books by teachers in France expose the fallacies of the popular 'child-centred' model of education. UK, London, An article in SPIKED, dated 28 April 2005, Retrieved from http://www.spiked-online.com/Articles/0000000CAAD6.htm, As visited on 20th Oct 2006 at 6:55 p.m.

Lewin, K. (1985). Quality in question: A new agenda for curriculum reform in developing countries. *Comparative Education, 21*(2), 117-133. Available from http://www.jstor.org.ezproxy.sussex.ac.uk/stable/3098967, As visited on 10 Jul 2010 at 9:05 p.m.

Lewin, K. (1997). Criterion referenced assessment: Panacea or palliative? In Bude, Udo, & Lewin, Keith (Eds.) *Improving Test Design: Constructing Test Instruments, Analysing Results and Improving Assessment Quality in Primary Schools in Africa* (Vol. 1). Bonn, Deutshe Stiftung fŸr internationale Entwicklung.

Linden, W. J. v. d. & Diao, Q. (2011). Automated Test-Form Generation. *Journal of Educational Measurement, 48*, 206–222. Retrieved online from: http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2011.00140.x/abstract, As visited on 29th Aug 2011 at 17:32 p.m.

Little, A. W. & Singh, J. S. (1992). Learning and Working: Elements of the Diploma Disease Thesis Examined in England and Malaysia. *Comparative Education, 28*(2), 181-200, Published by: Taylor & Francis, Ltd. Stable. Available from http://www.jstor.org/stable/3099430, As visited on 10th Jul 2010 at 09:40 p.m.

Mahajan, V. & Peterson, R. A. (1985). *Models for innovation diffusion*. Newbury Park, California, USA, SAGE Publications. Retrieved from http://books.google.com, As visited on 22nd Nov 2009 at 6:17 p.m.

Manaligod, H. J. T. (2009). Development of Prototype Software for Item Analysis and Item-Banking System for Likert Scales. *Alipato: A Journal of Basic Education, 3*(3), Via The U.P. Diliman Journals Online (UPDJOL). Retrieved online from: http://journals.upd.edu.ph/index.php/ali/article/viewFile/1759/1676, As visited on 10th Sep 2011 at 10:24 a.m.

Marso, R. N. & Pigge, F. L. (1987). Teacher-Made Tests and Testing: Classroom Resources, Guidelines, and Practices. Paper presented at the Annual Meeting of the Midwestern Educational Research Association (Chicago, IL, Oct 15-17, 1987). Retrieved online from: http://eric.ed.gov/PDFS/ED291781.pdf, As visited on 15th Aug 2011 at 10:35 a.m.

Maughan, P. & Willmott (2001). On-line Formative Assessment Item Banking and Learning Support. IN: Proceedings of the 5th CAA Conference, Loughborough: Loughborough University. Retrieved online from: https://dspace.lboro.ac.uk/dspace-jspui/handle/2134/1823, As visited on 31st Aug 2011 at 21:50 p.m.

Maxwell J. (2007). Types of generalization in qualitative research. USA, Columbia University, Teachers College. Retrieved online from: http://www.tcrecord.org/discussion.asp?i=3&aid=2&rid=12612&dtid=0&vdpid=2761, As visited on 12 Nov 2011 at 01:29 p.m.

McFee G. (1992). Triangulation in research: two confusions. *Educational Research*, *34*(3). Retrieved online from: , As visited on 27th Dec 2011 at 4:39 p.m.

McMillan, J. H. (1996). *Educational Research: Fundamentals for the consumer*. McMillan. - 2nd ed. HyperCollins Publishers Inc. Retrieved online from: http://www.odu.edu/~jritz/attachments/edrefu.pdf, As visited on 11 Nov 2011 at 03:40 p.m.

McNergney, R. F. & Carrier, C. A. (1980). *Teacher development*. New York, Macmillan.

McWilliam, Erica L (2002). Against Professional Development. *Educational Philosophy and Theory. 34*(3), 289-300. Retrieved online from: http://eprints.qut.edu.au/1032/1/erica2_against_professional_development_final.pdf, As visited on 13th Sep 2011 at 14:20 p.m.

Mergel, B. (1998). Instructional Design & Learning Theory. Retrieved from the website of College of Education, University of Saskatchewan, CANADA, URL: http://www.usask.ca/education/coursework/802papers/mergel/brenda.htm, As visited on 21st Nov 2005 at 14:40 a.m.

Miller, M. (1970). Computers can be used to score tests in order to reduce the workload of the teacher. Retrieved online from: http://eric.ed.gov, As visited on 15th Aug 2011 at 2:15 a.m.

Millman, J. & Westman, R. S. (1989). Computer-Assisted Writing of Achievement Test Items: Toward a Future Technology, *Journal of Educational Measurement, 26*(2), The Test Item, 177-190. National Council on Measurement in Education. Retrieved online from: http://www.jstor.org/stable/1434864, As visited on 30th Aug 2011 at 2:55 a.m.

Mitkov, R. & Ha, L. A.(2003). Computer-Aided Generation of Multiple-Choice Tests. UK, University of Wolverhampton, School of Humanities, Languages and Social Sciences. Retrieved online from: http://clg.wlv.ac.uk/papers/ruslan-NAACL-03.pdf, As visited on 28th Aug 2011 at 15:22 p.m.

Mitkov, R. Ha, L. A., Karamanis, N. (2005). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering 1*(1): 1–17. UK, c 200X Cambridge University Press. Retrieved online from: http://dl.acm.org/citation.cfm?id=1118897, As visited on 28th Aug 2011 at 14:55 p.m.

MoE web site (2005). The official web site of the Ministry of Education in the Kingdome of Saudi Arabia. Retrieved from http://www.moe.gov.sa/actualdata/general.aspx, As visited on 24th Sep 2005 at 17:35 a.m.

MoE website (2009). The official web site of the Ministry of Education in the Kingdome of Saudi Arabia, Retrieved from http://www.moe.gov.sa/, As visited on 13th Mar 2009 at 12:06 a.m.

Moon, J. (1999). *Reflection in learning & professional development: Theory and practice*. London, UK, RoutledgeFalmer. Retrieved from http://books.google.com, As visited on 14th Aug 2008 at 10:20 p.m.

Morrison, K. (1998). *Management theories for educational change*. London, UK, SAGE (Formerly: Paul Chapman). Retrieved from http://books.google.com, As visited on 5th Oct 2008 at 8:40 p.m.

Myron, H. (2005). Ask a scientist: Physics archive, NEWTON (an electronic community for Science, Math, and Computer Science K-12 Educators.), USA, Department of Energy, Argonne National Laboratory, Division of Educational Programs, Harold Myron, Ph.D., Division Director. Retrieved from http://www.newton.dep.anl.gov/askasci/phy00/phy00625.htm>, As visited on 16th June 2005 at 01:17 a.m.

Najjar, A. (2003). *Using SPSS in data analysis*. Riyadh, KSA, Data Net Establishment.

NASP (1997). Psychology applied to education: Lev S. Vygotsky's approach. Published in: *Communique (NASP), 25*(2), pp. 12-13. Retrieved online from the website of Centre for Cognitive-Developmental Assessment & Remediation, Part of: Psychological Services for Internationally Adopted Children, New York, USA, URL: http://www.bgcenter.com/Vygotsky_Appr.htm, As visited on 7th Sep 2011 at 19:40 p.m.

Nelson, L. (2002a). Lertap manual. Australia, Curtin University of Technology. Retrieved from http://lertap.curtin.edu.au/Documentation//Chapter1.doc., As visited on 15th Jul 2009 at 6:33 a.m.

Nelson, L. (2002b). Getting started with the LERTAP 5 prototype. Australia, Curtin University of Technolog. Retrieved from http://lertap.curtin.edu.au/Documentation/Archiveddocs.htm , As visited on 15th Jul 2009 at 6:50 a.m.

Nelson, L. (2005). Lertap electronic documentation. Australia, Curtin University of Technology. Retrieved from http://lertap.curtin.edu.au/History.htm, As visited on 15th Jul 2009 at 6:05 a.m.

Newman, J. M. (2000). Action research: A brief overview [14 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [On-line Journal], 1*(1). Retrieved from http://www.qualitative-research.net/fqs-texte/1-00/1-00newman-e.htm , As visited on 30 Dec 2004 at 12:23 a.m.

Newton, R. R. & Rudestam, K. E. (1999). *Your statistical consultant: Answers to your data analysis*. Sage Software Middle East. Retrived from http://books.google.com, As visited on 2nd Jan 2008 at 8:26 p.m.

Nitko, A. J. (1983). *Educational tests and measurements in education.* New York, London, Harcourt Brace Jovanovich.

Nitko, A. & Hsu, T. (1984a) Item Analysis Appropriate for Domain-Referenced Classroom Testing, A paper presented at the Annual Meeting of the American

Educational Research Association (68th, New Orleans, LA, April 23-27, 1984). Retrived online from: http://www.eric.ed.gov/PDFS/ED242781.pdf, As visited on 15 Aug 2011 at 10:25 a.m.

Nitko, A. & Hsu, T. (1984b) A Comprehensive Microcomputer System For Classroom Testing, *Journal of Educational Measurement, 21*(4), 377–390, December 1984, Retrieved Online from: http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1984.tb01041.x/abstract , As visited on 14 July 2011 at 22:37 p.m.

Noller, P. & Feeney, J. A. (2004). Studying family communication: Multiple methods and multiple sources. In Vangelisti A., Vangelisti A. L. (2003) *Handbook of family communication*. New Jersey, USA,, Lawrence Erlbaum Associates Inc. Retrieved from http://books.google.com, As visited on 3rd Dec 2009 at 11:49 p.m.

O'Donnell, Ed (2004). Use of forward versus backward reasoning during audit analytical procedures: Evidence from a computerised-process-tracing field study. *Accounting and Finance 44*, 75–95, Retrieved from http://www.people.ku.edu/~eod/research/manuscripts/study992.pdf, As visited on 28th Sep 2008 at 5:55 p.m.

Oakland, T. & Hambleton, R. K. (1995). *International perspectives on academic assessment*. USA, Kluwer Academic Publishers.

Obaidat, Th., Adas, A., Abdulhaq, K. (1996) *Research, Concept, tools, and Methods*. KSA, Riyadh, Dar Osama Publishers.

Olayyan, R. M. (2001). *Inquiry: Essentials, methodologies and procedures,* (In Arabic). Jordan, International Ideas Home Inc.

Oppenheimer, T. (1997) The Computer Delusion. The Atlantic Monthly, Digital Edition. Retrieved online from: http://www.theatlantic.com/issues/97jul/computer.htm, As visited on 12th Sep 2011 at 00:25 a.m.

Osborne, J. (2005). The nature of science and technology [PowerPoint slides]. Retrieved from http://cd.ed.gov.hk/sci/Lecture_e.htm, As visited on 21st Oct 2005 at 13:40 a.m.

Osterlind, S. J. (1992). *Constructing test items*. 2nd Ed., USA, Kluwer Academic Publishers.

Oxenham, J. (ed.) (1984). Education versus qualifications? A study of relationships between education, selection for employment and the productivity of labour.

Unwin Education Books series. London; Boston and Sydney: Allen & Unwin, Pp. x, 246. In Welfare Programs (1984). Consumer Economics; Urban and Regional Economics. *Journal of Economic Literature, 23*(2), 732-740. Published by: American Economic Association Available from http://www.jstor.org.ezproxy.sussex.ac.uk/stable/2725677, As visited on 10th Jul 2010 6:10 p.m.

Palya, W. L. (2000). The logical foundations of psychology. *Research Methods Lecture Notes,* Vol.1, Edition V2.2, USA, Jacksonville State University. JSU website, Retrieved from http://www.jsu.edu/depart/psychology/sebac/fac-sch/rm/Ch1-2.html#C-2, As visited on 28th Dec 2005 at 9:30 p.m.

Pearson Inc. (2005). *Comparing Standards-based Item Banks and Pre-built Tests for Classroom Assessment*. Pearson Education Inc. Retrieved online from: http://www.pearsonassessments.com/NR/rdonlyres/E4A8B6C7-6A62-4F52-9B68-B0498E5CC0B1/0/ItemBanks.pdf, As visited on 31st Aug 2011 at 23:30 p.m.

Pelgrum, W.J. (2001). Obstacles to the Integration of ICT in Education: Results from a Worldwide Educational Assessment. *Computers & Education 37*, 163–178. Retrieved online from: http://users.ntua.gr/vvesk/ictedu/article5_pelgrum.pdf, As visited on 3rd Sep 2011 at 16:20 p.m.

Penuel, W. R. & Yarnall, L. (2005). Designing handheld software to support classroom assessment: An analysis of conditions for teacher adoption. *Journal of Technology, Learning, and Assessment, 3*(5). Retrieved from http://www.jtla.org, As visited on 14th Mar 2009 at 9:08 p.m.

Postlethwaite, T. N. (2005). *Educational Research: Some Basic concepts and Terminology*. France, Paris, International Institute for Educational Planning, UNESCO. Retrieved online from: http://www.unesco.org/iiep/PDF/TR_Mods/Qu_Mod1.pdf, As visited on 15 Oct 2011 at 5:12 p.m.

Postman, N. (1992). *Technopoly: The Surrender of Culture to Technology*. USA, New York, Vintage Books. Retrieved online from: http://teach.boxwith.com/context/postman-technopoly.pdf, As visited on 13th Sep 2011 at 10:58 a.m.

Rabina, D. L. & Walczyk, D. J. (2007). Information professionals' attitude toward the adoption of innovations in everyday life. *Information Research, 12*(4). Retrieved

from http://informationr.net/ir/12-4/colis/colis12.html, As visited on 23rd Aug 2008 at 6:47 a.m.

Raskin, J. D. (2002). Constructivism in psychology: Personal construct psychology, radical constructivism, and social constructionism, *American Communication Journal, 5*(3). Retrieved from http://www.acjournal.org/holdings/vol5/iss3/special/raskin.htm, As visited on 20th Oct 2006 at 2:30 a.m.

Rezq, H. (2008) Effect of Implementing Constructivist Learning in Teaching Maths on Academic Achievement of Pupils of the First Year Intermediate School at Makkah City. (Phd Thesis). KSA, University of Umm Al-Qura. Retrieved online from: http://libback.uqu.edu.sa/hipres/FUTXT/5981.pdf, As visited on 1st Oct 2011 at 3:39 p.m.

Riyadh Daily Newspaper (2002, Mar 24th). Front page, Riyadh, KSA, Al-Yamamah Journalism Corporation.

Riyadh Daily Newspaper (2002, Mar 25th). The electronic edition, Retrieved from http://www.alriyadh.com/Contents/2002/03/25-03-2002/page12.html.

Riyadh Daily Newspaper (2002, Mar 26th). The electronic edition, Retrieved from http://www.alriyadh.com/Contents/2002/03/26-03-2002/page12.html.

Robinson, W. & Austin, W. (1969). A Computerized test-correcting service for teachers. California Educational Research Association, Retrieved online from: http://eric.ed.gov/PDFS/ED027753.pdf, As visited on 15th Aug 2011 at 11:07 a.m.

Roger, C. (1999). A Primer in Diffusion of Innovations Theory. Australia, Australian National University. © Xamax Consultancy Pty Ltd, Online Notes, Recent revised 26 September 1999. Retrieved online from: www.anu.edu.au/people/Roger.Clarke/SOS/InnDiff.html, As visited on 23rd Aug 2008 at 5:22 a.m.

Rondinelli, D. A., Middleton, J. & Verspoor, A. M. (1990). *Planning education reforms in developing countries: The contingency approach*. Durham and London. UK, Duke University Press.

Ross, K. (2005) *Sample Design for Educational Survey Research*. France, International Institute for Educational Planning/UNESCO, Retrieved online from: http://www.iiep.unesco.org/fileadmin/user_upload/Cap_Dev_Training/Training_Materials/Quality/Qu_Mod3.pdf, As visited on 11 Nov 2011 at 3:46 p.m.

Russell, T. & Munby, H. (1992). *Teachers and teaching: From classroom to reflection.* London, UK, The Falmer Press.

Salih, Ahmed, Zaki (1988). *Educational psychology.* In Arabic, 10th Ed., Egypt, Al Ma'aref Publishers.

Satterly, D. (1994). *Quality in external assessment*, In Harlen, W. (1994). *Enhancing quality in assessment*. London, Paul Chapman Publishing Ltd.

School Material (2001). Thinking science package. UK, Monifieth High School, Angus Council.

Shaukat, A., Arain, T. M., Alam, M. F., & Shahid, A. (2007). Teaching strategies and academic performances of undergraduates in Quaid-i-azam *Journal of College of Physicians & Surgeons Pakistan, 17*(10), 598-602, Pakistan, bahawalpur, Medical College. Retrieved from http://www.cpsp.edu.pk/jcpsp/ARCHIEVE/JCPSP-2007/oct07/article5.pdf, As visited on 29th Aug 2008 at 1:14 a.m.

Shayer, M. (2000). *GCSE 1999: Added-value from schools adopting the CASE Intervention*. London, UK, The Centre for Advancement of Thinking, The School of Education, King's College, London.

Shepard, L. A. (2000). *The role of classroom assessment in teaching and learning.* USA, The Regents of the University of California. Retrieved from http://www.cse.ucla.edu/products/Reports/TECH517.pdf, As visited on 13th Mar 2009 at 9:18 p.m.

Shook, C. A. (2004). Promoting thinking through pedagogical changes in science lessons, *Asia-Pacific Forum on Science Learning and Teaching, 5*(3), Foreword (Dec., 2004). Retrieved from http://www.ied.edu.hk/apfslt/v5_issue3/foreword/index.htm, As visited on 27th Oct 2006 at 3:42 p.m.

Shuttleworth, M. (2008). Case Study Research Design. Retrieved online from: http://www.experiment-resources.com/case-study-research-design.html, As visited on 22nd Oct 2011 at 8:09 p.m.

Silye M. F. & Wiwczaroski T. B. (2002) A Critical Review of Selected Computer Assisted Language Testing Instruments. Retrieved online from: www.date.hu/acta-agraria/2002-01i/fekete1.pdf, As visited on 6th Sep 2011 at 11:52 p.m.

Simons, H. (2009). Case Study Research in Practice, USA, SAGE. Retrieved online from: http://books.google.com, As visited on 16th Nov 2011 at 8:57 a.m.

Smawley, R. B. (1962). Teacher-made Tests Improved by Use of Shortcut Item Analysis. Taylor & Francis, Ltd. Retrieved online from: http://www.jstor.org.ezproxy.sussex.ac.uk/stable/pdfplus/30194115.pdf?acceptTC=true, As visited on 16th Aug at 13:12 a.m.

Smits, H., Wang, H., Towers, J., Crichton, S., Field, J., & Tarr, P. (2005). Deepening understanding of inquiry teaching and learning with E-portfolios in a teacher preparation program. *Canadian Journal of Learning and Technology, 1*(3) Fall. Retrieved from http://www.cjlt.ca/content/vol31.3/smits.html, As visited on 23rd Aug 2008 at 11:44 p.m.

Somekh, B., Lewin, C. (2004). *Research Methods in the Social Sciences*. SAGE Publications. Retrieved online from: http://lib.myilibrary.com.ezproxy.sussex.ac.uk?ID=36883, As visited on 24th Sep 2011 at 13:12 p.m.

Souchon, A. & Lings, I. (2008). *Adopting internal marketing practices across national borders: Key propositions and implications*. Retrieved from http://smib.vuw.ac.nz:8081/WWW/ANZMAC2001/anzmac/AUTHORS/pdfs/Souchon.pdf, As visited on 5th Oct 2008 at 2:12 a.m.

Soy, S.K. (1997). The case study as a research method. Unpublished paper, USA, University of Texas at Austin. Retrieved online from: http://www.gslis.utexas.edu/~ssoy/usesusers/l391d1b.htm, As visited on 22nd Oct 2011 at 7:45 p.m.

SPSS (2001). SPSS for Windows, Release 11.0.0 standard version, SPSS Inc.

Stage, C. (1998). A comparison between item analysis based on item response theory and classical test theory: A study of the SweSAT subtest WORD. Sweden, Dept of Educational Measurement, Faculty of Social Sciences, Umeå University. Retrieved from http://www8.umu.se/edmeas/publikationer/index_eng.html, As visited on 12th Jul 2002 at 7:02 p.m.

Stake, R. E. (1995). The art of case study research. USA, SAGE. Retrieved online from: http://books.google.com, As visited on 16th Nov 2011 at 8:51 a.m.

Stanley, J. C. & Hopkins, K. D. (1978). *Educational and psychological measurement and evaluation.* New Delhi. India, Prentice Hall of India Private limited.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning, *Kappan Professional Journal* [Online Journal]. Last updated 6 June 2002, Phi Delta Kappa International. Retrieved from http://electronicportfolios.org/afl/Stiggins-AssessmentCrisis.pdf, As visited on 27th Aug 2008 at 8:55 p.m.

Student Assessment Document (2000). First edition, KSA, Ministry of Education.

Sukamolson S. (Not-dated). Computerized Test/Item Banking and Computerized Adaptive Testing for Teachers and Lecturers. Thailand, Chulalongkorn University, Language Institute. Retrieved online from: http://www.stc.arts.chula.ac.th/ITUA/Papers_for_ITUA_Proceedings/Suphat2.pdf, As visited on 31st Aug 2011 at 23:07 p.m.

Sultana, Q. (2001). Scholarly teaching -Application of bloom's taxonomy in Kentucky's classrooms, ERIC Database, Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/27/f7/fd.pdf, As visited on 3rd Sep 2005 at 2:57 a.m.

SurveyMethods (2009). Website of SurveyMethods, Inc. Retrived online from: http://blog.surveymethods.com/3-types-of-probability-samples-for-educational-survey-research/, As visited on 12 Nov 2011 At 12:39 p.m.

Soydan, H. (1998) Evaluation research and social work, *International Journal of Social Welfare*, 7(2), 74–78, April 1998, First published online 3 Apr 2007, Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1468-2397.1998.tb00205.x/abstract, As visited on 2nd Dec 2010 at 11:00 a.m.

Tashkandi, M. O. (1981). Effects of instruction and personal traits of Saudi pre-service science teachers on the use of higher cognitive questions (Doctoral dissertation). An Arbor, Michigan, USA, Indiana University, University Microfilms International.

Teaching & Learning Resource Centre (2002). *Using item analysis reports to improve the quality of your test questions*. Carlton University website, Canada, Retrievd from URL: http://www.carleton.ca/tlrc/scantron/analysis.htm#item, As visited on 20th Jul 2002 at 10:55 p.m.

TESOL (2003). Qualitative Research: Case Study Guidelines. *TESOL Quarterly, 37*(1), 157-178. Retrieved online from: http://www.tesol.org/s_tesol/sec_document.asp?cid=476&did=2153, As visited on 18th Oct 2011 at 4:20 p.m.

Theory Cluster (2004). Diffusion of innovations theory, Notes of University of Twente, Last modified on 09/06/2004 15:11:50, © University of Twente. Retrieved from http://www.tcw.utwente.nl/theorieenoverzicht/Theory%20clusters/Communication%20and%20Information%20Technology/Diffusion_of_Innovations_Theory.doc, As visited on 23rd Aug 2008 at 11:38 p.m.

Thurmond V. A. (2001). The Point of Triangulation. *Journal Of Nursing Scholarship, 33*(3), 253-258. Retrieved online from: http://www.ruralhealth.utas.edu.au/gr/resources/docs/the-point-of-triangulation.pdf, As visited on 27th Dec 2011 at 5:12 p.m.

Trochim, W. (2006). Research Methods Knowledge Base, Retrieved from <http://www.socialresearchmethods.net/kb/intreval.php >, As visited on 2nd Dec 2010 at 09:50 a.m.

Value Based Management.net (2004). A management portal specifically aimed at the information needs of senior executives with an interest in value creation, managing for value and valuation. Retrieved from http://www.valuebasedmanagement.net/methods_rogers_innovation_adoption_curve.html, As visited on 23rd Aug 2008 at 3:43 a.m.

Vaughan, W. (2002). Professional development and the adoption and implementation of new innovations: Do teacher concerns matter? *The International Electronic Journal for Leadership in Learning (IEJLL), 6*(5), University of Calgary Press. Retrieved from http://www.ucalgary.ca/iejll/vaughan, As visited on 8th July 2010 at 12:25 p.m.

Vrasidas, C., & Glass, G. V. (2004). *Online professional development for teachers: The impact of teacher education*. Charlotte, USA, Center for the Application of Information Technologies, Information Age Publishing (IAP). Retrieved from http://books.google.com, As visited on 23 Aug 2008 at 10:53 p.m.

Wainer, H. (1989). The Future of Item Analysis. *Journal of Educational Measurement, 26*(2), 191-208. National Council on Measurement in Education. Retrieved online from: http://www.jstor.org/stable/1434865, As visited on: 17th Dec 2010 at 05:42 a.m.

Walpole, R. E. (1976). *Elementary statistical concepts*. New York, USA, Macmillan Publishing co. Inc.

Wang T. H., Wang, K. H., Wang, W. L., Huang, S. C., and Chen, S. Y. (2004) Web-based Assessment and Test Analyses (WATA) System: Development and

Evaluation, *Journal of Computer Assisted Learning, 20*(1), 59–71, February 2004, Article first published online: 3 FEB 2004, Retrieved online from http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2729.2004.00066.x/full,As visited on 4th Aug 2011 at 00:28 a.m.

Wang, T. H., Wang, K. H., & Huang, S. C. (2008). Designing a Web-based Assessment Environment for Improving Pre-service Teacher Assessment Literacy. *Computers & Education 51*, 448–462, Retrieved online from: www.sicencedirect.com , As visited on 4th Aug 2011 at 00:44 a.m.

Warnick B. R. (2001) A review of the Book: Let Them Eat Data: How Computers Affect Education, Cultural Diversity, and the Prospects of Ecological Sustainability. By Bowers, C. A. (2000). USA, Athens, The University of Georgia Press. The review is retrieved online from: , As visited on 12th sep 2011 at 23:34 p.m.

Weiss D. (2011). *Item Banking, Test Development, and Test Delivery*. In Press, The APA Handbook on Testing and Assessment. Washington DC, American Psychological Association, Retrieved online from: http://www.assess.com/docs/Weiss_Handbook_Chapter.pdf, As visited on 31st Aug 2011 at 21:35 p.m.

Wellington, J. (2000). *Teaching and Learning Secondary Science: Contemporary issues and practical approaches*. London, UK, Routledge.

White, Y. B. & Frederiksen J. R. (1998). Inquiry, modelling, and metacognition: Making science accessible to all students. *Cognition and instruction, 16*(1), 3-118, Lawrence Erlbaum Associates, Inc. Retrieved from http://131.193.130.213/media//white_b_etal_1998.pdf, As visited on 11th Nov 2009 at 9:18 a.m.

Wikipedia (2009). The free encyclopaedia that anyone can edit. Retrieved from http://en.wikipedia.org/wiki/File:InnovationLifeCycle.jpg, and http://en.wikipedia.org/wiki/File:Diffusionofideas.PNG, As visited on 20th May 2009 at 12:30 p.m. See note 14.

Williams, J. M. (1991). Writing quality teacher-made tests: A handbook for teachers [ERIC database]. Retrieved from http://www.eric.ed.gov, As visited on 4th Sep 2005 at 23:18 p.m.

Winch, C. & Forman-Peck, L. (2000). Teacher professionalism, educational aims and action research: The evolution of policy in the United Kingdom. *Teacher Development, 4*(2).

Winefield, H. R. (2003). *Occupational Stress in the Service Professions. Psychological Aspects*, Retrieved from http://books.google.com, As visited on 22nd Mar 2008 at 11:15 p.m.

Wolfe, D., Overbaugh, R., & Bol, L. (2005). *A teaching methodology/content survey, introduction of the survey: Theoretical background*, Norfolk, Darden College of Education, Old Dominion University. Retrieved from http://www.odu.edu/educ/idt/eci_survey/survey_preamble.html, As visited on 25th Aug 2005 at 18:23 p.m.

Wright B. D. & Bell, S. B. (1984). Item Banks: What, Why, How. *Journal of Educational Measurement, 21*(4), Application of Computers to Educational Measurement, 331-345. National Council on Measurement in Education. Retrieved online from: http://www.jstor.org/stable/1434585, As visited on 15th Aug 2011 at 10:11 p.m.

WTLRCCU. (2002) Website of Teaching and Learning Resource Centre of Carlton University. Retrieved from http://www.carleton.ca/tlrc/scantron/s_intro.htm, As visited on 29th Jul 2002 at 8:35 p.m.

Wayne, N. (1976), A Computer Support System for a Teacher Evaluation Model. Corporate source: Beaverton School District 48, OR., pp. 16. Retrieved online from: http://www.datastarweb.com.ezproxy.sussex.ac.uk/EDUCAT/20110911_154335_bc123_2/WBDoc12/5001/aa10773a/, As visited on 11th Sep 2011 at 17:02 p.m.

Yang, J., Han, X., & Zhou Q. (2011). The Design and Development of a Semi-Auto Computer Generated Testing Paper System — A Case Study in the School of Continuing Education at China University Of Geosciences. *TOJET: the Turkish Online Journal of Educational Technology, 10*(2). Retrieved online from: http://www.eric.ed.gov/PDFS/EJ932240.pdf, As visited on 29th Aug 2011 at 18:07 p.m.

Yin, R. K. (1994*). Case Study Research, Design and Methods*. 2nd edition. Thousand Oaks: Sage. Retrieved online from:

http://www.soberit.hut.fi/~mmantyla/work/Research_Methods/Case_Study/Case%20Study%20Research.doc. As visited on 18th Oct 2011 at 4:40 p.m.

Yu, Chong-ho (Alex) (2002). Which reliability coefficients should I use? Contact the author at: Josephine Wai-chi Wong, PO Box 612, Arizona State University, Tempe AZ 85280, Email: asumain@yahoo.com.hk, Source: The homepage of Dr. Chong-ho Yu (Alex). Retrieved from http://seamonkey.ed.asu.edu/alex/teaching/assessment/alpha.html, As visited on 20th Jul 2002 at 10:55 p.m.

Yu, Chong-ho (Alex) (2006) Using SAS for Classical Item Analysis and Option Analysis. Contact the author at: Josephine Wai-chi Wong, PO Box 612, Arizona State University, Tempe AZ 85280, Email: asumain@yahoo.com.hk, Source: The homepage of Dr. Chong-ho Yu (Alex). Retrieved from http://seamonkey.ed.asu.edu/~alex/, As visited on 14th Mar 2006

Zahranee, S. S. (1998). The extent to which preparatory school science teachers in makkah are competent on test construction skills and their level of practice in this respect. (Masters dissertation). Makkah, KSA, Um Al-Qura University.

Appendices

Appendix 1 - Section 1

Employee Evaluation Forms

Appendix Table 1.1: Former performance evaluation criteria

Ser	Item	Score
1	Adherent to use eloquent Arabic.	6
2	Keen to organise and undertake school activities: both those out of curriculum and related ones	5
3	Give attention to develop his knowledge	5
4	Attending work on time	7
5	Skilful educationally in preparing and providing lessons	7
6	His/her ability of the subject matter and ability of fulfilling its aims	7
7	Giving attention to on-going assessment and individuals' differences	7
8	Planning the curricular contents and the level of congruency of what have been applied so far with time.	4
9	The use of the board, textbooks and other instructional aids	4
10	Skilful on providing lessons and keeping classroom discipline	5
11	Level of pupils academic achievement	10
12	Giving attention to use and mark Homework and worksheets.	5
13	General behaviour (being good example)	4
14	Appreciating responsibility	4
15	Accepting guidance	4
16	Showing wisdom in due situations	4
17	Relationship with his/her chairperson	4
18	Relationship with colleagues	4
19	Relationship with pupils and their parents	4

Appendix Table 1.2: Recent performance evaluation criteria

		Items	Maximum Grades			Given Grade
			Category (A)	Category (B)	Category (C)	
A – Performance Evaluation	A	1. The ability to make sound decisions.	5			
	AB	2. Skill of planning & achieving the goals.	5	6		
	AB	3. Full knowledge of the job policies & procedures.	7	7		
	AB	4. Awareness of the integration between education & knowledge.	5	6		
	AB	5. Ability to develop job styles.	5	5		
	AB	1. Skill of directing & following-up	7	8		
	AB	7. Taking care of school environment & making use of it.	6	5		
	ABC	8. Using classic Arabic.	6	6	6	
	ABC	9. Sharing in school activities.	5	4	5	
	ABC	10. Capacity & desire to broaden perspective & increase professional development.	5	5	5	
	ABC	11. Compiles with establishing work hours	7	7	7	
	ABC	12. Proficiency of the subject & the ability to achieve its goals.	5	7	7	
	ABC	13. Awareness of the educational basics of preparation & application.		6	7	
	C	14. Considering continuous evaluation & individual differences.			7	
	C	15. Distribution of the syllabus according to the schedule.			4	
	C	16. Using the board, textbooks & other learning aids.			4	
	C	17. Presentation skills & class management.			5	
	C	18. Level of students' achievements.			10	
	C	19. Application & Home assignments are graded promptly.			5	
	TOTAL					

B- Personality	AB	20. The ability to lead a discussion.		4	4	
	ABC	21. General Behaviour. (a role model)	4	4	4	
	ABC	22. Being Responsible	4	4	4	
	ABC	23. Accepting Directions	4	4	4	
	ABC	24. Good Manners	4	4	4	
	TOTAL					

C- Relations	ABC	25. with superiors	4	4	4	
	ABC	26. with co-workers	4	4	4	
	AC	27. with parents & students	4		4	
	TOTAL					

Category (A): for the headmistress, Category (B): for the supervisor, and trainers Category (C): for the teachers

Appendix 1 - Section 2

Review of Some Item Analysis Software Packages

Some common similar software packages illustrated

There are a number of software packages that perform test item analysis and lie under CTT or IRT; in this illustration I am going to present a review for some common ones under CTT since the CAIAT software package relates to this stream. I will then explain why the present research's CAIAT represents a better alternative to be adopted for this research purpose by highlighting its major characteristics and main functions.

ITEMAN© is one of the most well-known software packages for IAT and has been used for decades since the PCs were working under MS DOS© system. However one of the observed disadvantages of ITEMAN© is that although its developers have upgraded it into MS Windows© and hence designed its menus and interface into a Windows© compatible versions, they maintained the input file as the former method that contains a stack of lines of code (LOC) as Appendix Figure 1.1 shows. This LOC should follow a specific order that is determined by ITEMAN© system, thus the user needs to study this in order to be able to comply with it. For example, the first line is called the control line that describes initial information about the data that will be entered and a half page paragraph of ITEMAN's manual is to explain how to deal with it accordingly, as Appendix Figure 1.2 illustrates. The second line includes the key for items answers, the third line is for the number of alternatives (since ITEMAN© is confined to analyse multiple-choice tests only), the fourth is to combine each item to the corresponding scale and then the following lines contain examinees' answers. Instead, this could be made easier for the user by designing screens of data entry in a graphical way that guide the user on how to enter each bit of data without having to understand or memorise the original ITEMAN© data entry scheme which I already presented.

Appendix Figure 1.1: ITEMAN©'s data file for entering data (ASC, 2006)

```
45 S N 10
ABDABDCACBCDCABCDBDABDCACBCDDBABCADCBDABDBDAD KEY
44444444444444444444444444444444444444444444 N ALTS
00000000000000012211221122112211221122112211 SCALES
ID#998783 ABDACBBDDDAACCBDCACDCBCCBCDCDACDACCACBBCBBBAD
ID#989367 AADCABBDCDACBBCCBDAABCCCACDBSSSAAACACDBBBDDCN
ID#103102 ABDACBADCBAACBBDCACBBDCAAABBDABDCDCBCDCBDDDAA
ID#993167 ABDACBADCBDACBBDCACBCDCABCDBDABDCACDCDBBDBCAD
ID#983458 CADBDAABADBBBBCAACDDACBBADBDCCAAABBADCCCBANNN
.
.
.
```

Items on the first scale are assigned inclusion codes of "0", items on the second

Appendix Figure 1.2: ITEMAN©'s manual: first line explained (ASC, 2006).

The Control Line

The first line of the data file must contain the following data in the columns specified:

Column Data

1-3 Number of items for which responses are recorded for each examinee (maximum is 750)
4 Blank
5 Alphanumeric code for omitted responses
6 Blank
7 Alphanumeric code for items not reached by the examinee
8 Blank
9-10 Number of characters of identification data recorded for each examinee (maximum is 80)

In columns 1-3, you *must* enter the number of items that are included in the file. *This number must be right-justified:* The "units" go into column 3, the "tens" in column 2, and the "hundreds" in column 1. A maximum of 750 items can be handled by ITEMAN Figure 2-1 shows a data file with 30 items to be analyzed.

Column 5 must contain the alphanumeric code for items that the examinee has omitted. This can be a digit larger than the number of alternatives, a letter, or some other character including a "blank." For example, it might be "9" for a five-alternative item, an "O" for omitted, or a period. Column 7 must contain the alphanumeric code for items that the examinee did not reach and therefore did not have a chance to answer. Like the omission code, it can be a digit larger than the number of alternatives or any other character. In Figure 2-1, the letter "**O**" indicates an omitted item, and "**N**" indicates a not-reached item.

For dichotomously scored items, ITEMAN makes no distinction between items that are omitted and items that are not reached; *both are scored as incorrect*. For rating-scale-type items, ITEMAN will allow you to specify one of three ways for dealing with missing data (these options are presented on the Options screen). Both the responses recorded as omitted and those recorded as not reached will be considered "missing."

Columns 9 and 10 contain the number of characters at the beginning of each examinee's data record used for identification. As with the number of items, *these digits must be right justified* — the "tens" must be in column 9 and the "units" in column 10. The maximum number of identification characters is 80. If columns 9 and 10 are left blank or if zero identification characters are specified, examinee identification will not be expected and the examinees' responses must begin in column 1 on the data lines. The example in Figure 2-1 indicates that there are 5 characters of identification for each examinee; in the data lines (beginning on line 5 of the input file in Figure 2-1), you will note that the examinees are identified by "`EX001`" through "`EX005`"

Appendix Figure 1.3 illustrates an output file of ITEMAN© in which *P* values appear under "Prop. Correct" column and *D* values under the "Disc. Index" column. The Alternative statistics section serves as a way for judging power of distractors by

referencing statistics of low and high groups and "Prop. Total" ratio. This report appears in a design similar to old fashion designs of programs under MS DOS© system where there are no lines or boxes such as those that appear in recent years' reports. I think that ITEMAN© developers might have kept old fashion design of their screens and reports to provide their former users of the similar shapes of their original product which their customers are used to.

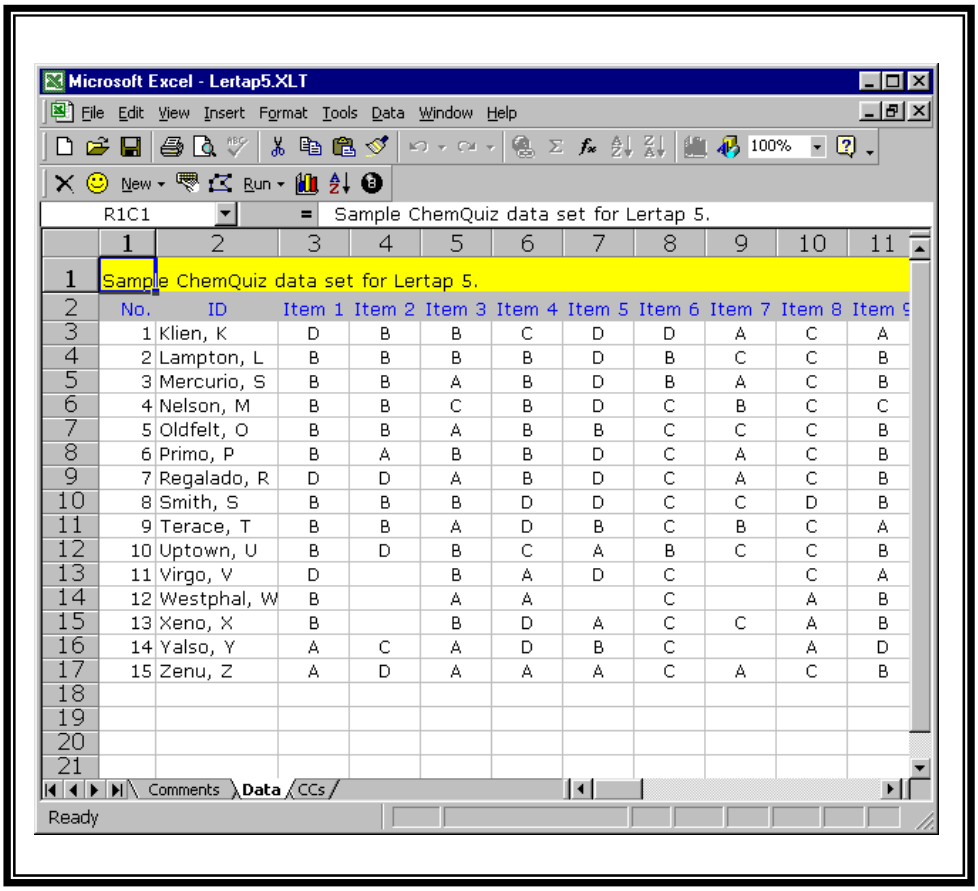
Appendix Figure 1.3: ITEMAN©'s output file for analysis results (ASC, 2006)

Item Statistics					Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	1-1	.58	.63	.37	1	.16	.33	.10	-.45	
					2	.58	.22	.85	.37	*
					3	.08	.06	.00	-.16	
					4	.18	.39	.05	-.50	
					Other	.00	.00	.00		
2	1-2	.38	.53	.38	1	.26	.44	.10	-.50	
					2	.14	.06	.00	-.21	
					3	.38	.17	.70	.38	*
					4	.22	.33	.20	-.39	
					Other	.00	.00	.00		
3	1-3	.20	.04	-.13	1	.24	.44	.15	-.49	
					2	.20	.11	.15	-.13	*
CHECK THE KEY					3	.16	.22	.15	-.28	
2 was specified, 4 works better					4	.40	.22	.55	.15	?
					Other	.00	.00	.00		

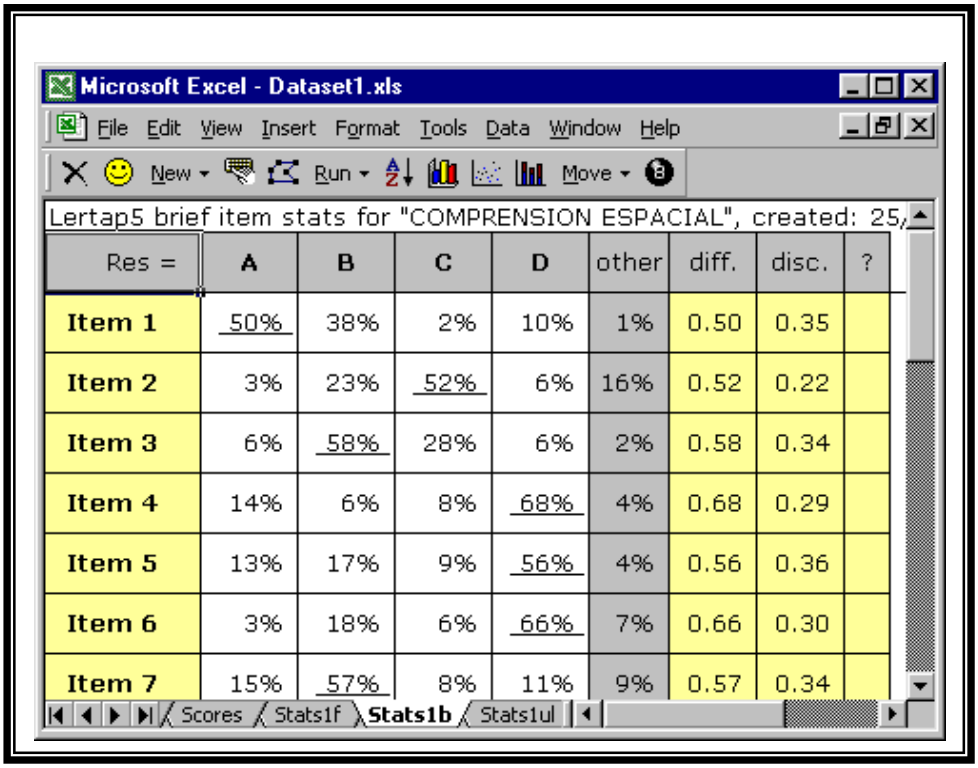
The Laboratory of Educational Research Test Analysis Package (Lertap©) is another software that has been designed in Canada 1972, developed through different programming languages and used in different countries in America, Europe and Asia until recently designed to work under the shell of Excel© software[48] at Curtin University of Technology, Australia (Nelson 2002a and Nelson 2005).

Unlike ITEMAN©, Lertap© utilises Excel's interface for data entry (Appendix Figure 1.4) which is represented by conventional worksheets that any beginner user of Excel© is accustomed to. Output reports on the other hand, as the example that appears on Appendix Figure 1.5, come in graphical coloured and outlined shapes that aid the user to distinguish boundaries of printed figures easily.

Appendix Figure 1.4: Example of Lertap© sheet for data entry (Nelson 2002a)



Appendix Figure 1.5: Example of Lertap© output sheet (Nelson 2002b)



Finally, I present Appendix Table 1.3, which compares some common CTT and IRT software packages to give a glance idea about this aspect.

Appendix Table 1.3: A List of software packages available by ASC under CTT or IRT framework (ASC, 2009)

Product	Platform	Maximum Items	Maximum Examinees	Non-Academic Price	Academic Price
Classical Item Analysis					
Integrity	Online Application	500	Varies	Inquire	Inquire
ITEMAN	Windows	750	Unlimited	$299	$299
Lertap 5	Windows/Macintosh	255	65,535	$399	$199
Lertap 5	Excel 2007 (Windows)	Over 10,000	Over 1,000,000	(Requires current license).	(Requires current license).
Scrutiny.	Windows	1,000	Unlimited	$399	$299
TestFACT	Windows	1,000	Unlimited	Inquire	$250
Rasch Analysis (1-parameter IRT)					
Quest	DOS/Macintosh	400	10,000	$460	$460
RASCAL	Windows	750	Unlimited	$299	$299
RSP	DOS	96	Unlimited	$405	$280
RUMMFOLD*ss*	Windows	100	5,000	$400	$400
RUMMFOLD*pp*	Windows	30	5,000	$400	$400
WINMIRA	Windows			$500	$200
2- and 3-Parameter IRT Analysis					
BILOG-MG	Windows	1,000	Unlimited	Inquire	$250
LOGIMO	DOS			$405	$280
MSP	Windows	100	32,000	$625	$480
MULTILOG	Windows	Unlimited	Unlimited	Inquire	$250
PARELLA	DOS	60	300	$405	$280
PARSCALE	Windows	Unlimited	Unlimited	Inquire	$250
XCALIBRE	Windows	750	Unlimited	$399	$399

The CAIAT Package Illustrated

The CAIAT software is a computer program which carries out IAT statistical calculations electronically. It is written in Visual Basic® language, works under Windows® for IBM® compatible PCs and has an Arabic language interface. It uses driven menus and dialogue box forms for interacting with the user, while printing its reports both on screen and printer using tabulated forms that include summaries and

explanations. The user enters raw data of test items' marks by means of different ways: all items for each pupil or all pupils for each item; then runs calculations of test analyses by no more than a click of a button. This output could be seen on screen or be printed out in forms that are clear, simple, and understandable to teachers, principals, and supervisors. Resulting analyses indicators in these reports include: difficulty coefficient *P*, discrimination index *D* and reliability index for each test item. These could be calculated for both types of questions: objective and essay questions and there is a special processing of statistical calculations related to multiple-choice items by calculating distraction efficiency index for every alternative.

The software is characterised by its ease of use and ergonomic advantage. In fact, every user who exhibits basic computer skills and understands IAT concepts can use it with minimal training; or maybe without. It comes with three users' manuals in Arabic: "Installation Manual" which explains how to install the software the first time, "User Manual" which explains its different screens, functions, reports. Helpful tips are provided, while the "Reference Manual" explains IAT concepts from the educational assessment perspective. The software also has the facility to store the entered data on databases that can be retrieved later either to be edited, deleted, re-analysed, or reprinted. Because of this feature, it could be installed on the school's PC and then used by all teachers where each can store his or her data on an independent file without interference from others.

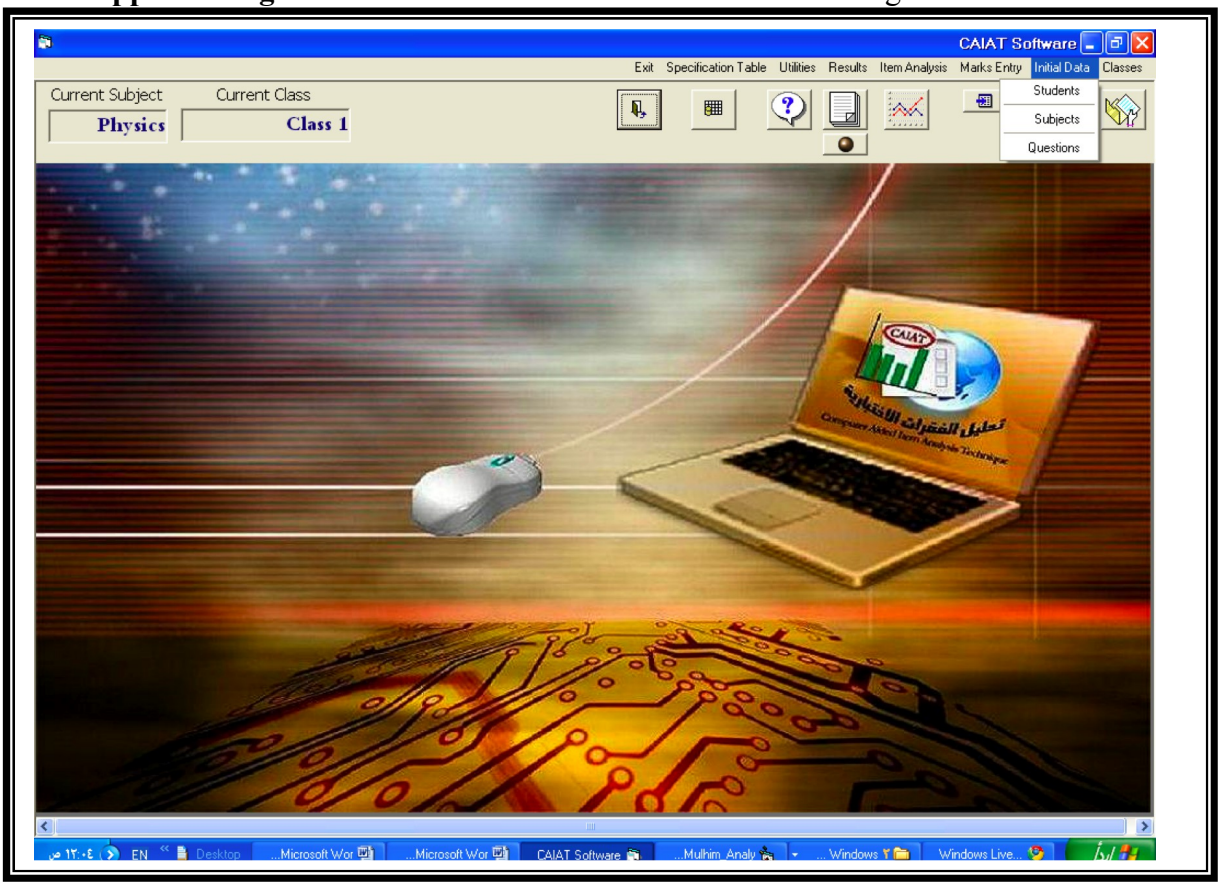
As an advanced programmer using MS Visual Basic©, I have developed this software package in its entirety[49]. I have designed its structure, screens and reports. I have written the code of the program that does the calculations of item analysis parameters *P* and *D* and the other statistical parameters. I followed up its application to recognise the required amendments needed to improve its performance and acted upon them in order to reach an optimal level that meets requirements of facility and ease of use that this research encourages. This product is dedicated for this research and is its main technical component. The CAIAT package is a Windows© based software and comes on a form similar to that used by many popular software packages such as MS Word© or Power Point©, where it uses scroll down menus supported by shortcut buttons. The contents of these menus are logically identical to the same flow of tasks for hand calculations of IAT. This makes learning CAIAT congruent with learning IAT steps. The CAIAT software provides a help function that illustrates how to use the software using plenty of pictures and explaining in good detail, as appropriate, how to deal with

some situations and what to do for some critical functions. It also provides some alerts of anticipated user errors, adding to its functionality as a comprehensive guide.

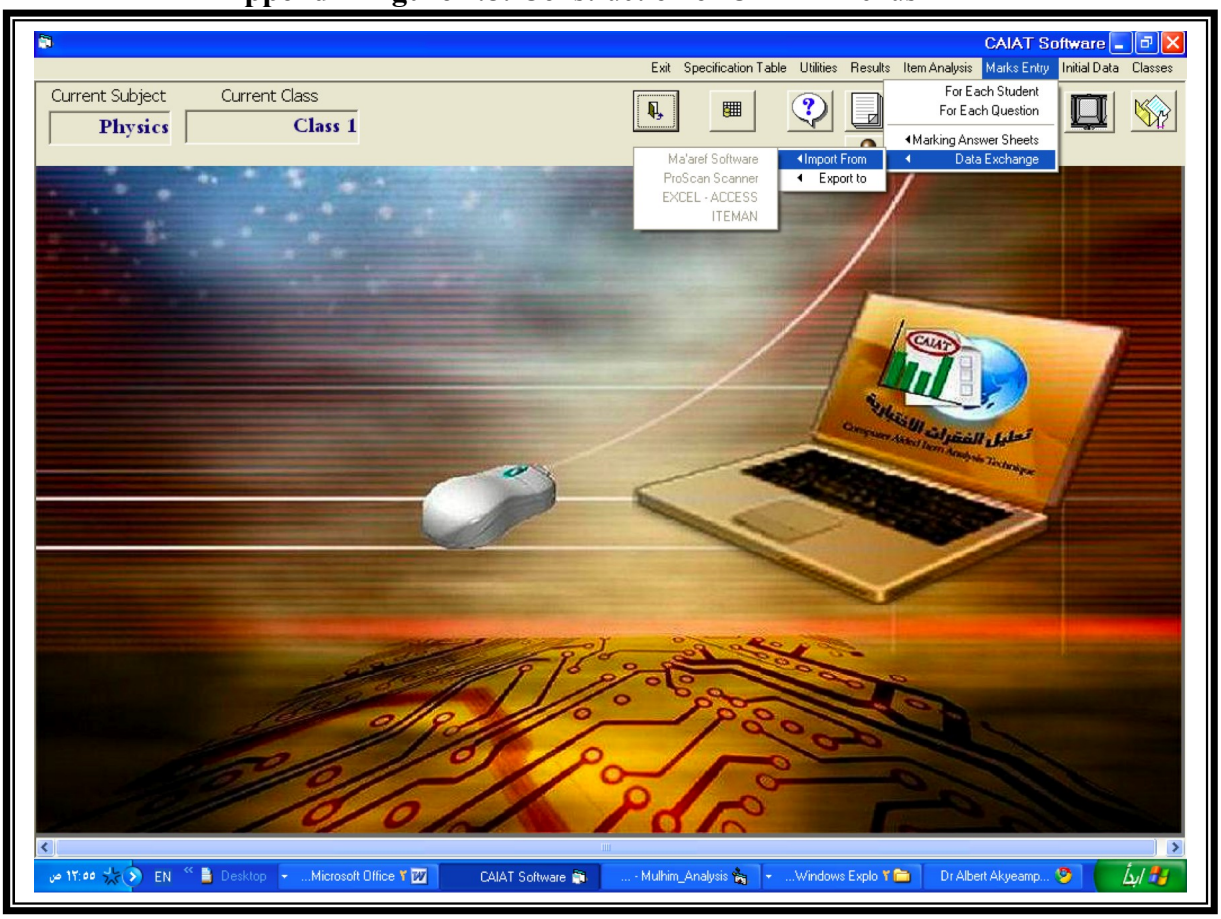
Appendix Figure 1.6 shows the first screen of this software. Since the CAIAT software is in Arabic, I have translated its screens in English just for the purpose of this research report; therefore Appendix Figure 1.6 is repeated in Appendix Figure 1.7 as an English version. However, diagrams that follow will be in English. Appendix Figure 1.8 shows how the menus of the CAIAT software are constructed. Appendix Figure 1.9 shows how marks are entered, Appendix Figure 1.10 shows how you enter one mark and the software repeats it for all pupils, This occurs because there are cases where the majority of pupils answer an item correctly, thereby earning a similar high mark. The software repeats that mark for all pupils. The user should then go through the data and update only those data for the pupils that did not earn the entered mark. This method gives a chance to the user to enter data quickly if this situation applies (it might apply for the mark zero as well). Appendix Figure 1.11 shows a screen for setting up initial data for an item. These are type of item (objective/subjective), number of alternatives, correct alternative and mark of the item. I have to indicate that by assigning here that the item is subjective, formulas for calculating *P* and *D* that correspond to subjective items (essay questions for example) will be applied whilst by assigning here that the item is objective other formulas for objective items will be applied. These formulas have been explained in the CTT section of Chapter 4. This characteristic is not mentioned in any other software package amongst the packages that I have encountered and examined during my search, which gives the CAIAT software additional advantage. Appendix Figure 1.12 shows outcomes of the software in a report that gives item analysis results and Appendix Figure 1.13 shows a report that gives power of distractors. The aim of presenting these screens is to show the way CAIAT graphical screens facilitate the user data entry process, which represents a major aspect of the program. Furthermore, its reports appear in these diagrams in modern graphical tabulated forms.

Appendix Figure 1.6: First screen of the CAIAT software Arabic version

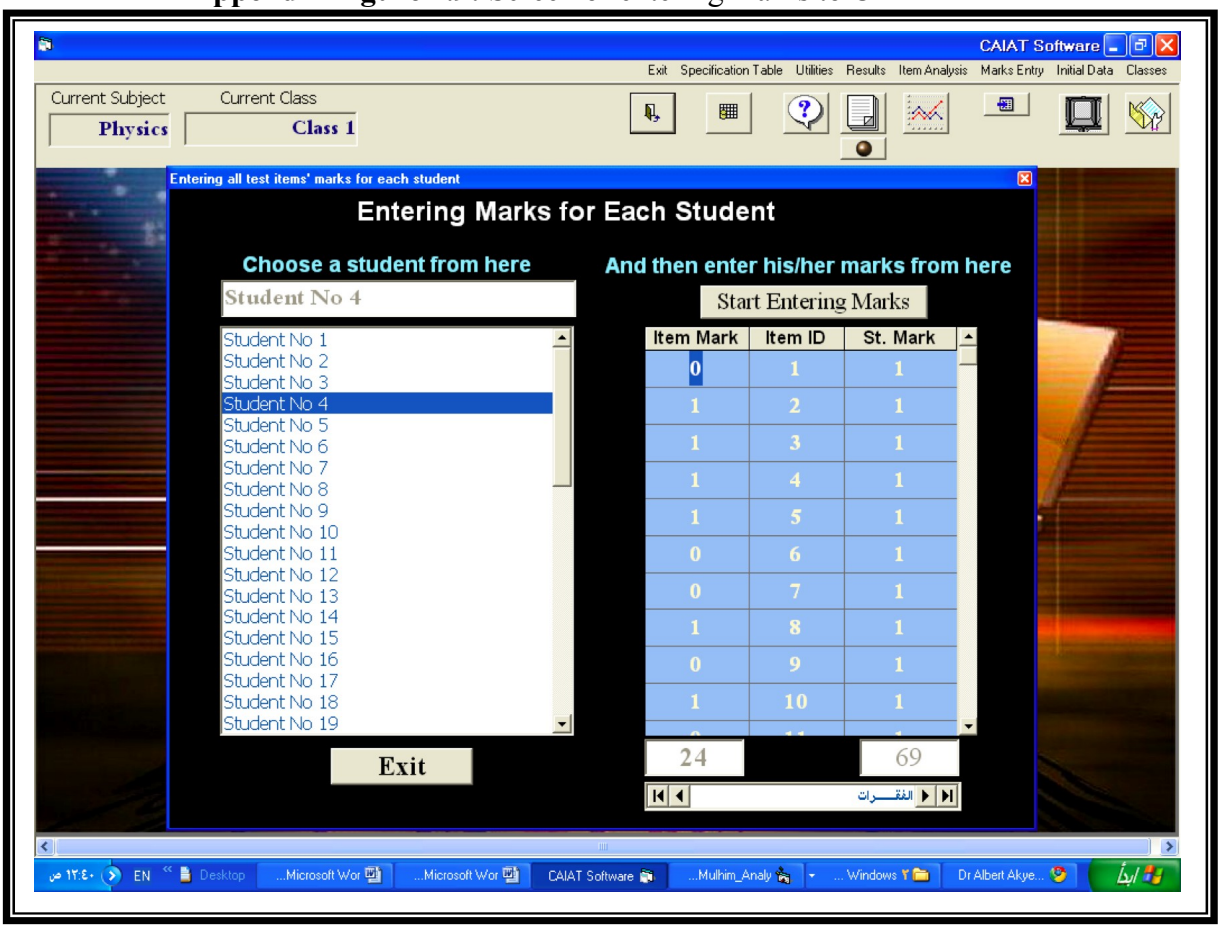
Appendix Figure 1.7: First screen of the CAIAT software English version



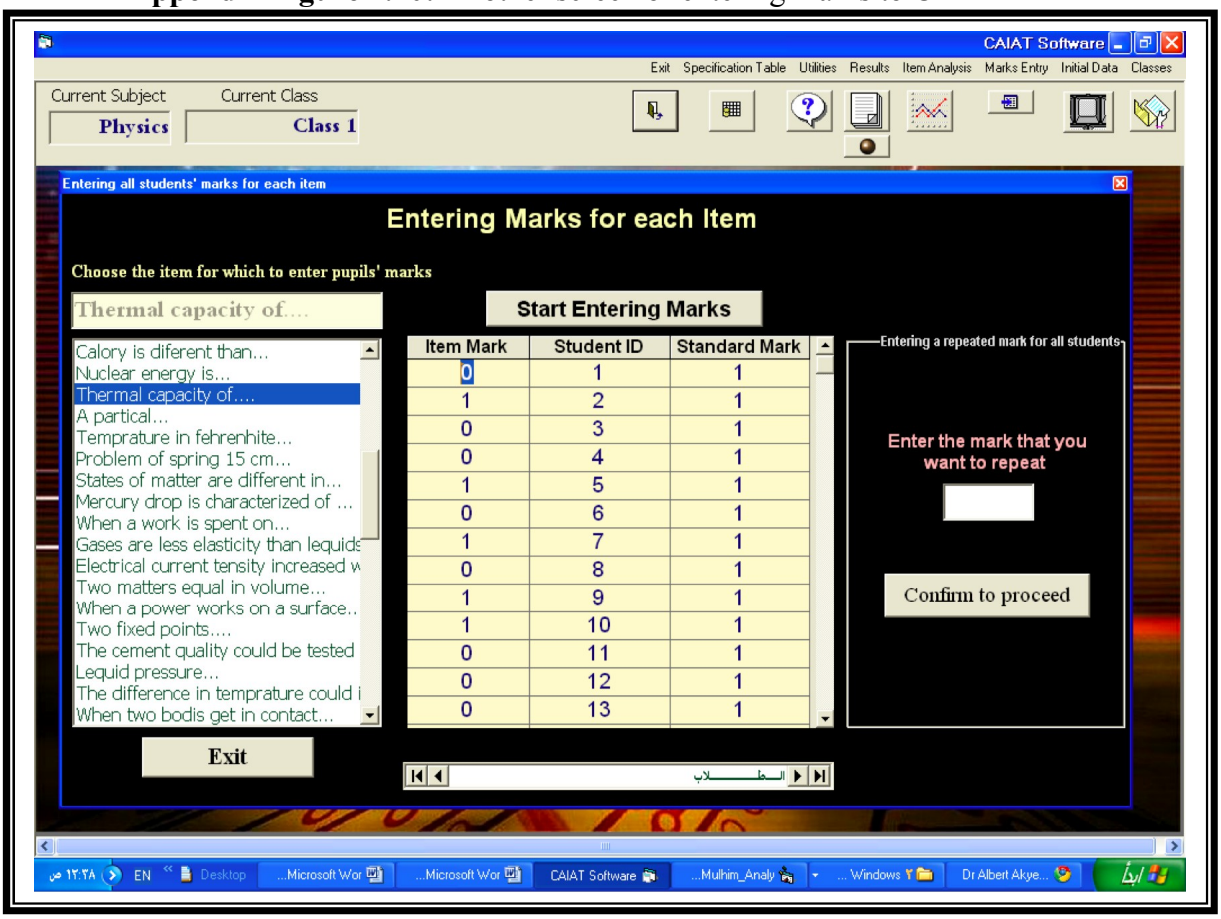
Appendix Figure 1.8: Construction of CAIAT menus



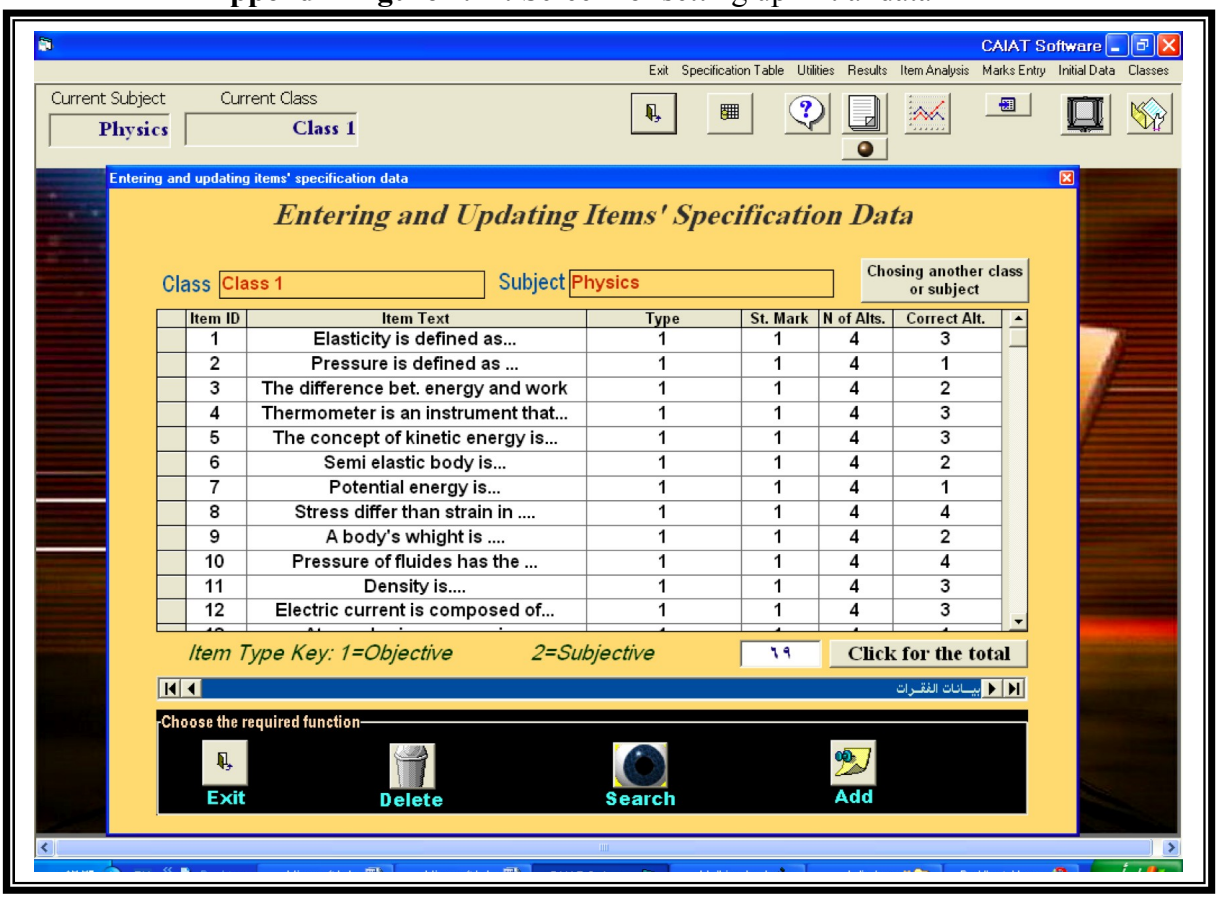
Appendix Figure 1.9: Screen of entering marks to CAIAT



Appendix Figure 1.10: Another screen of entering marks to CAIAT



Appendix Figure 1.11: Screen for setting up initial data



Appendix Figure 1.12: Report for analysis results

CAIAT Software Package

Kingdome of Saudi Arabia

Ministry Of Education

Items Analysis Results

Item Reliability	Discrimination Index	Difficulty Coeffecient	Text of the Question	Item Code
79%	-4%	82%	Elasticity is defined as…	1
65%	28%	74%	Pressure is defined as …	2
73%	4%	82%	The difference bet. energy and work	3
68%	28%	74%	Thermometer is an instrument that…	4
63%	32%	72%	The concept of kinetic energy is…	5
55%	24%	64%	Semi elastic body is…	6
44%	8%	60%	Potential energy is…	7
33%	32%	52%	Stress differ than strain in ….	8
39%	28%	54%	A body's whight is ….	9
60%	20%	70%	Pressure of fluides has the …	10
44%	-4%	58%	Density is….	11
52%	32%	64%	Electric current is composed of…	12
39%	12%	54%	Atmospheric pressure is…	13
47%	8%	60%	Hook's law requires…	14
60%	20%	70%	Temprature is…	15
33%	36%	50%	Derivative physical quantity is…	16
25%	20%	42%	Quantity of heat has the…	17
15%	20%	18%	Newton is…	18
39%	16%	12%	Calory is diferent than…	19
23%	36%	34%	Nuclear energy is…	20
15%	0%	20%	Thermal capacity of….	21
44%	-8%	16%	A partical…	22
33%	0%	12%	Temprature in fehrenhite…	23
41%	24%	56%	Problem of spring 15 cm…	24
52%	32%	64%	States of matter are different in…	25
36%	24%	52%	Mercury drop is characterized of …	26
23%	4%	42%	When a work is spent on…	27
60%	20%	70%	Gases are less elasticity than lequids	28
52%	24%	64%	Electrical current tensity increased	29
15%	44%	34%	Two matters equal in volume…	30
23%	24%	24%	When a power works on a surface…	31
25%	-20%	30%	Two fixed points….	32
15%	24%	36%	The cement quality could be tested	33
23%	12%	22%	Lequid pressure…	34
17%	20%	38%	The difference in temprature could	35
20%	16%	40%	When two bodis get in contact…	36
20%	52%	38%	The following quantities	37
23%	-4%	42%	Two cubes of iron…	38

CAIAT Software 1 \ 2 30 رجب، ١٤٣١

Appendix Figure 1.13: Report for power of distractors

CAIAT Software

Kingdome of Saudi Arabia

Ministry of Education

Distractors Effectiveness Report

7th Alt.	6th Alt.	5th Alt	4th Alt.	3rd Alt.	2nd Alt.	1st Alt.	Item's efffectivns	Corect Alt.	Item Rel.	Discr. Index	Diff. Coeff.	N. Of Alt.	Item Code
0%	0%	0%	0%	84%	8%	8%	Hi Group	3	79%	-4%	82%	4	1
0%	0%	0%	0%	84%	0%	16%	Low Group		Elasticity is defined as...				
0%	0%	0%	0%	0%	-8%	8%	D. Effeciency						
0%	0%	0%	0%	4%	4%	92%	Hi Group	1	65%	28%	74%	4	2
0%	0%	0%	12%	24%	8%	56%	Low Group		Pressure is defined as ...				
0%	0%	0%	12%	20%	4%	36%	D. Effeciency						
0%	0%	4%	0%	4%	80%	12%	Hi Group	2	73%	4%	82%	4	3
0%	0%	0%	0%	8%	80%	12%	Low Group		The difference bet. energy and work				
0%	0%	-4%	0%	4%	0%	0%	D. Effeciency						
0%	0%	0%	0%	92%	0%	8%	Hi Group	3	68%	28%	74%	4	4
0%	0%	0%	8%	60%	0%	32%	Low Group		Thermometer is an instrument that...				
0%	0%	0%	8%	32%	0%	24%	D. Effeciency						
0%	0%	0%	0%	88%	4%	8%	Hi Group	3	63%	32%	72%	4	5
0%	0%	0%	8%	56%	20%	16%	Low Group		The concept of kinetic energy is...				
0%	0%	0%	8%	32%	16%	8%	D. Effeciency						
0%	0%	0%	8%	8%	80%	4%	Hi Group	2	55%	24%	64%	4	6
0%	0%	0%	28%	12%	52%	8%	Low Group		Semi elastic body is...				
0%	0%	0%	20%	4%	28%	4%	D. Effeciency						
0%	0%	0%	32%	4%	0%	64%	Hi Group	1	44%	8%	60%	4	7
0%	0%	0%	28%	4%	16%	52%	Low Group		Potential energy is...				
0%	0%	0%	-4%	0%	16%	12%	D. Effeciency						
0%	0%	0%	68%	8%	16%	8%	Hi Group	4	33%	32%	52%	4	8
0%	0%	0%	32%	12%	36%	20%	Low Group		Stress differ than strain in				
0%	0%	0%	36%	4%	20%	12%	D. Effeciency						
0%	0%	0%	4%	20%	68%	8%	Hi Group	2	39%	28%	54%	4	9
0%	0%	0%	8%	8%	40%	44%	Low Group		A body's whight is				
0%	0%	0%	4%	-12%	28%	36%	D. Effeciency						
0%	0%	0%	80%	16%	4%	0%	Hi Group	4	60%	20%	70%	4	10
0%	0%	0%	60%	16%	8%	16%	Low Group		Pressure of fluides has the ...				
0%	0%	0%	20%	0%	4%	16%	D. Effeciency						
0%	0%	0%	4%	56%	12%	28%	Hi Group	3	44%	-4%	58%	4	11
0%	0%	0%	4%	60%	8%	28%	Low Group		Density is....				
0%	0%	0%	0%	-4%	-4%	0%	D. Effeciency						
0%	0%	0%	0%	80%	4%	16%	Hi Group	3	52%	32%	64%	4	12
0%	0%	0%	4%	48%	0%	48%	Low Group		Electric current is composed of...				
0%	0%	0%	4%	32%	-4%	32%	D. Effeciency						
0%	0%	0%	16%	8%	16%	60%	Hi Group	1	39%	12%	54%	4	13
0%	0%	0%	20%	28%	4%	48%	Low Group		Atmospheric pressure is...				
0%	0%	0%	4%	20%	-12%	12%	D. Effeciency						
0%	0%	0%	4%	12%	64%	20%	Hi Group	2	47%	8%	60%	4	14
0%	0%	0%	8%	8%	56%	28%	Low Group		Hook's law requires...				
0%	0%	0%	4%	-4%	8%	8%	D. Effeciency						
0%	0%	0%	80%	4%	8%	8%	Hi Group	4	60%	20%	70%	4	15
0%	0%	0%	60%	20%	8%	12%	Low Group		Temprature is...				
0%	0%	0%	20%	16%	0%	4%	D. Effeciency						
0%	0%	0%	4%	68%	20%	8%	Hi Group	3	33%	36%	50%	4	16
0%	0%	0%	8%	32%	28%	32%	Low Group		Derivative physical quantity is...				
0%	0%	0%	4%	36%	8%	24%	D. Effeciency						
0%	0%	0%	32%	4%	52%	12%	Hi Group	2	25%	20%	42%	4	17
0%	0%	0%	56%	4%	32%	8%	Low Group		Quantity of heat has the...				
0%	0%	0%	24%	0%	20%	-4%	D. Effeciency						
0%	0%	0%	20%	28%	24%	28%	Hi Group	1	15%	20%	18%	4	18
0%	0%	0%	32%	44%	16%	8%	Low Group		Newton is...				
0%	0%	0%	12%	16%	-8%	20%	D. Effeciency						
0%	0%	0%	60%	8%	20%	12%	Hi Group	2	39%	16%	12%	4	19
0%	0%	0%	48%	28%	4%	20%	Low Group		Calory is diferent than...				
0%	0%	0%	-12%	20%	16%	8%	D. Effeciency						

Appendix 2

The Questionnaire and Pre-test

This research instrument is used at the beginning of the training course where it aims to provide information about the trainees' background prior to training. All of the sample teachers should fill in this instrument.

.----------.

Questionnaire/Pre-test

(This instrument is to be delivered to teachers at the beginning of the introduction stage).

Dear teacher,

This Questionnaire/Pre-test is for research purposes only and will not be reflected in your evaluation or otherwise. Its objective is to measure the extent to which you may have any prior fundamentals and/or skills in using computers, utilising higher-order cognitive concepts, and/or analysing tests' questions. Therefore, we would appreciate an accurate appraisal of your actual skills level in the interest of collecting valid data for our research.

(Optional**) Name:** ..... **School:** .....

Trainee No.: .....

Graduated from college of education? ( ) Yes ( ) No

Your level of graduation: ( ) Excellent ( ) Very good ( ) Good ( ) Satisfactory

No. of Years of Experience: ( ) 5 years or less ( ) 6 - 10 years ( ) Greater than 10 years

==========

Questionnaire

1 – When writing your tests, do you make a specification table? **( ) Yes ( ) No**

2 – Have you ever attended a training course on test construction? **( ) Yes ( ) No**

If your answer is yes, when was that? Was it before a period of:

( ) 6 months ( ) 6 months – 1year ( ) 1 - 2 years ( ) 3-4 years ( ) more than 4 years

3 – Have you ever attended a training course on test results analysis?

If your answer is yes, when was that? Was it before a period of:

( ) 6 months ( ) 6 months – 1year ( ) 1 - 2 years ( ) 3-4 years ( ) more than 4 years

4 – Considering the test results, have you ever noticed a problem in any of your test questions? **( ) Yes ( ) No**

5 - Do you know how to use computers? **( ) Yes ( ) No**

If your answer is No then skip the following questions and jump to section 2 (Pre-test):

6 – Have you ever installed any software package on a PC? **( ) Yes ( ) No**

7 –Tick all the software applications that you can use from the list below:

- Microsoft Word®. - Microsoft Explorer®. - Microsoft Excel®.
- Microsoft PowerPoint®. - Microsoft Access®. - Microsoft Outlook®.
- SPSS®.

8- In the table below, please describe the extent to which you are using a computer for each purpose:

Computer Purpose	**All the time**	**Mostly**	**From time to time**	**Rarely**	**I Don't use**
As a word processor.					
To manage my personal financial affairs.					
For presenting (lessons, training courses, etc.)					
For scheduling tasks, appointments, etc.					
Browsing and searching the Internet.					
Doing calculations (other than personal).					
Searching databases for personal purposes (such as telephone directories, names or addresses of people, organisations or products, etc.)					
Searching databases for non-personal purposes (such as electronic encyclopaedias, dictionaries and reference books, etc.)					

Pre-test: Section 1

If you find that you have no idea about any of the following questions then please do not answer any of them and tick the following statement and then go to section 3:

◯ **I do not know anything about the topic of these questions**

The following three questions are multiple-choice questions based on one correct answer, so circle only the identifier of the correct answer (in case of changing your mind, cross the cancelled one).

1- Bloom's Taxonomy is:
- A- A directory of lesson plans' levels.
- B- A curriculum design methodology.
- C- A classified levels of cognition.
- D- A teaching aids' categorization scheme.

2 – Bloom's Taxonomy consists of:
- A – 3 levels.
- B – 4 levels.
- C – 5 levels.
- D – 6 levels.

3 – The highest level of cognitive demand of Bloom taxonomy is:
- A – Comprehension.
- B – Understanding.
- C – Mastery.
- D – Evaluation.

The following questions are True/False questions, so tick one of the last two columns:

T e x t	True	False
4 – The first higher cognitive level is recall of information.		
5 – Explaining a concept is one aspect of fulfilling application level.		
6 – Application level of cognition requires that the learner be able to combine two facts or concepts to form a new one.		
7 – Analysis level of cognition means that the learner can single out the components that form a fact or a concept.		
8 – Teaching on higher cognitive levels requires that the teacher participates much more than the learner.		
9 – As guidance, there are common verbs suitable for each level of cognition in Bloom's Taxonomy.		
10 – Teaching thinking (or meta thinking) is one tool for teaching on higher cognitive levels.		

Questions 11-13 require that for each fact mentioned, you write down four questions, each of which is for a specific level of cognition as shown in the table below:

Q11. ***The Fact***	***Light scatters when it is directed to a non-smooth opaque surface. This is because such surfaces, on a micro scale, have ups and downs that cause some incident light to be scattered or diffusely reflected off the surface in many directions.***
Recall	
Comprehension	
Application	
Analysis	
Q12. ***The Fact***	***Transporting electricity over a very long distance requires using cables of a substantial great thickness. This is because a substance's resistance of electricity (R) depends proportionally on the length of the transporting cable (L) and adversely on the area of that cable's cross section (A).*** *R = C (L/A) where C: Constant*
Recall	
Comprehension	
Application	
Analysis	
Q13. ***The Fact***	***Charles' law represents the proportional relationship between temperature and volume of gas that is under a constant pressure.***
Recall	
Comprehension	
Application	
Analysis	

Pre-test: Section 2

If you find that you have no idea about any of the following questions then please do not answer any of them and tick the following statement:

◯ **I do not know anything about the topic of these questions.**

The following three questions are multiple-choice questions based on one correct answer, so circle only the identifier of the correct answer (in case of changing your mind, cross the cancelled one).

1- Item Analysis Technique is a method of knowing what is wrong with the:
- A- Lesson plans of the teacher.
- B- Activities of instruction for a specific lesson.
- C- The questions teachers ask by a written test.
- D- The behavioural objectives of a lesson.

2 – Item Analysis Technique requires:
- A – Moderate statistical ability.
- B – Good background of planning.
- C – Good selection of items.
- D – Technical skills in instruction.

3 – In Item Analysis Technique, we calculate discrimination index because a good test:
- A –should discriminate amongst pupils.
- B –should not discriminate amongst pupils.
- C –should avoid any type of racial discrimination.
- D –should discriminate from other tests.

The following questions are True/False questions, so tick one of the last two columns:

T e x t	True	False
4 – To explore the level of ambiguity of a test question we need to measure its difficulty coefficient.		
5 – To explore to what extent a question is efficient in showing pupils' individual differences, we need to calculate its discriminating index.		
6 – For a given question, if its difficulty coefficient value is 95% and its discrimination coefficient is 5% then this indicates that it is a good question.		
7 – Item analysis technique aids us to judge quality of testing items upon pupils' responses to those items rather than the subjective opinion of an individual who will be looking at that test.		
8 – Discrimination index may appear in negative values.		
9 – If discrimination index value is near zero then this is an indicator to a good question.		
10 – The optimum value of difficulty coefficient is 0.5.		

Model Answers

Section 1

1.C 2. D 3. D 4. False 5. False 6. False

7. True 8. False 9. True 10. True

.	
Q11. The Fact	***Light scatters when it is directed to a non-smooth opaque surface. This is because such surfaces, on a micro scale, have ups and downs that cause some incident light to be scattered or diffusely reflected off the surface in many directions.***
Recall	*What happens when light hit a non-smooth opaque surface?*
Comprehension	*What do we mean by the word "scatter" when we say that light scatters when it hits a non-smooth opaque surface?*
Application	*How can you make use of the fact that light scatters when it hits a non-smooth opaque surface?*
Analysis	*Demonstrate what happens to the components of light when it hits a non-smooth opaque surface?*
Q12. The Fact	***Transporting electricity over a very long distance requires using cables of a substantial great thickness. This is because a substance's resistance of electricity (R) depends proportionally on the length of the transporting cable (L) and adversely on the area of that cable's cross section (A).*** *R = C (L/A) where C: Constant*
Recall	*What is the characteristic factor of a cable which you should take in account when transporting electricity over a very long distance?*
Comprehension	*When transporting electricity over a very long distance we have to use a substantially thick cable… Why?*
Application	*Give an example that shows some function of the formula R = C (L/A). Your example should be other than transporting electricity over very long distances.*
Analysis	*Explain on a micro scale why the speed of electricity increases with thicker cables?*
Q13. The Fact	***Charles' law represents the proportional relationship between temperature and volume of gas that is under a constant pressure.***
Recall	*Write down Charles' law?*
Comprehension	*If the pressure is not constant in Charles law what would be the resulting effect on the concept that underlies the law?*
Application	*How can you make use of Charles law to increase the upward flowing speed of a balloon?*
Analysis	*Describe what happens to the molecules of a gas that is under Charles law?*

Section 2

1.C 2. A 3. A 4. True 5. True 6. False 7. True 8. True 9. False 10. True

Appendix 3

The Training Course Syllabus

Introduction to Computers Session (Day 0)

This session will be dedicated to those who have no idea or enough practical experience about computers. Although it is a one day course, it is going to be given one week before the next set of sessions. This is to allow enough time for more self-practice and/or (if needed) more training on using computers skills.

Objectives:

1. *To have a general idea about what is computer, how does it work and what it is primarily used for.*
2. *To be able to use mouse and keyboard.*
3. *To be able to use the main functions and services of Windows such as running a software, file and folder concepts, opening and closing folders, the icon concept, the desktop concept and shutting down.*
4. *To be able to use WORD fundamentally in terms of how to launch WORD, type, delete, edit, print on screen, and handle the printer.*

Contents

1. Introduction to computers.
2. Discover computer parts and initial Windows® screen.
3. How to use the mouse.
4. How to run a software package (WORD® as an example).
5. How to use the keyboard in WORD®: Writing text, deleting and editing.
6. More on how to use the mouse: application: how to print on screen and by printer.
7. How to deal with printers: switching on and off, feeding with papers, dealing with jammed paper and taking care of printers.
8. How to close WORD®.
9. What is the meaning of Folder, File, Icon and Desktop; including training on how to deal with each item.
10. More hands-on training through challenging tasks.

HCD Session (Day1)

Objectives:

1. *To learn Bloom's Taxonomy and behavioural objectives and be able to write down a HCD instructional behavioural objectives successfully.*
2. *To discriminate among different types of questions and understand each type's criteria for optimum test.*
3. *To be able to construct good HCD questions.*

Contents

1. Bloom's Taxonomy.
2. Behavioural Objectives.
3. Types of Questions.
4. Optimum Questions Criteria.
5. Small Project.

IAT Session (Day2)

Objectives:

1. *To understand the meaning of difficulty and discrimination from IAT point of view.*
2. *To recognize the statistical background of difficulty coefficient and discrimination index calculations.*
3. *To be able to judge test items upon IAT concepts.*

Contents

1. Difficulty Coefficient (D).
2. Discrimination Index.
3. Hands-on Examples.
4. Small Project.

CAIAT Session (Day3)

Objectives:

1. *To be able to install the CAIAT software.*
2. *To be able to use the CAIAT software in data entry, reports and setup.*
3. *To be able to judge test items upon IAT concept using the CAIAT software.*

Contents

1. Installing CAIAT Software.
2. Types of Data Entry.
3. Reports.
4. Setup.
5. Small Project.

Appendix 4

Post Test

This research instrument is used at the end of the last day of training where it aims to provide information about the trainees' skills that they achieved by training. It is administered to all of the sample teachers.

.----------.

Dear teacher,

This Questionnaire/Pre-test is for research purposes only and will not be reflected in your evaluation or otherwise. Its objective is to measure the extent to which you may have any prior fundamentals and/or skills in using computers, higher-order cognitive concepts and/or analysing tests' questions. Therefore, we would appreciate an accurate appraisal of your actual skills level in the interest of collecting valid data for our research.

(Optional**) Name:** ..... **School:** .....

Trainee No.: .....

Graduated from college of education? ( ) Yes ( ) No

Your level of graduation: ( ) Excellent ( ) Very good ( ) Good ( ) Satisfactory

No. of Years of Experience: ( ) 5 years or less ( ) 6 - 10 years ( ) Greater than 10 years

==========

Post-test: Section 1

If you find that you have no idea about any of the following questions then please do not answer any of them and tick the following statement and then go to section 3:

◯ I do not know anything about the topic of these questions

The following three questions are multiple-choice questions based on one correct answer, so circle only the identifier of the correct answer (in case of changing your mind cross the cancelled one).

1- Bloom's Taxonomy is:

A- A directory of lesson plans' levels.
B- A curriculum design methodology.
C- A classified levels of cognition.
D- A teaching aids' categorization scheme.

2 – Bloom's Taxonomy consists of:

A – 3 levels.
B – 4 levels.
C – 5 levels.
D – 6 levels.

3 – The highest level of cognitive demand of Bloom taxonomy is:

A – Comprehension.
B – Understanding.
C – Mastery.
D – Evaluation.

The following questions are True/False questions, so tick one of the last two columns:

T e x t	True	False
4 – The first higher cognitive level is recall of information.		
5 – Explaining a concept is one aspect of fulfilling application level.		
6 – Application level of cognition requires that the learner be able to combine two facts or concepts to form a new one.		
7 – Analysis level of cognition means that the learner can single out the components that form a fact or a concept.		
8 – Teaching on higher cognitive levels requires that the teacher participates much more than the learner.		
9 – As guidance, there are common verbs suitable for each level of cognition in Bloom's Taxonomy.		
10 – Teaching thinking (or meta thinking) is one tool for teaching on higher cognitive levels.		

Questions 11-13 require that for each fact mentioned, you write down four questions, each of which is for a specific level of cognition as shown in the table below:

Q11. ***The Fact***	***Light scatters when it is directed to a non-smooth opaque surface. This is because such surfaces, on a micro scale, have ups and downs that cause some incident light to be scattered or diffusely reflected off the surface in many directions.***
Recall	
Comprehension	
Application	
Analysis	
Q12. ***The Fact***	***Transporting electricity over a very long distance requires using cables of a substantial great thickness. This is because substance's resistance of electricity (R) depends proportionally on the length of the transporting cable (L) and adversely on the area of that cable's cross section (A).*** *R = C (L/A) where C: Constant*
Recall	
Comprehension	
Application	
Analysis	
Q13. ***The Fact***	***Charles' law represents the proportional relationship between temperature and volume of gas that is under a constant pressure.***
Recall	
Comprehension	
Application	
Analysis	

Post-test: Section 2

The following three questions are multiple-choice questions based on one correct answer, so circle only the identifier of the correct answer (in case of changing your mind cross the cancelled one).

1- Item Analysis Technique is a method of knowing what is wrong with the:
- A- Lesson plans of the teacher.
- B- Activities of instruction for a specific lesson.
- C- The questions teachers ask by a written test.
- D- The behavioural objectives of a lesson.

2 – Item Analysis Technique requires:
- A – Moderate statistical ability.
- B – Good background of planning.
- C – Good selection of items.
- D – Technical skills in instruction.

3 – In Item Analysis Technique we calculate discrimination index because a good test:
- A –should discriminate amongst pupils.
- B –should not discriminate amongst pupils.
- C –should avoid any type of racial discrimination.
- D –should discriminate from other tests.

The following questions are True/False questions, so tick one of the last two columns:

T e x t	True	False
4 – To explore the level of ambiguity of a test question we need to measure its difficulty coefficient.		
5 – To explore to what extent a question is efficient in showing pupils' individual differences we need to calculate its discriminating index.		
6 – For a given question, if its difficulty coefficient value is 95% and its discrimination coefficient is 5% then this indicates that it is a good question.		
7 – Item analysis technique aids us to judge quality of testing items upon pupils' responses to those items rather than the subjective opinion of an individual who will be looking at that test.		
8 – Discrimination index may appear in negative values.		
9 – If discrimination index value is near zero then this is an indicator to a good question.		
10 – The optimum value of difficulty coefficient is 0.5.		

Appendix 5

Questionnaire to Teachers After the Contingent Application Stage

This research instrument is administered to all of the sample teachers at the beginning of the workshop where it aims to provide information about the trainees' initial response to what they have trained for.

.----------.

Dear teacher,

This Questionnaire is for research purposes only and will not be reflected in your evaluation or otherwise. Its objective is to measure the extent to which you may have tried out yourself in applying what you have learnt in the training course of the HCD/CAIAT project. Therefore, we would appreciate an accurate appraisal of your actual practice in the interest of collecting valid data for our research.

(Optional**) Name:** ..... **School:** .....

Trainee No.: ..... (This is the number that you have chosen as an identifier when you filled down the first questionnaire at the training session)

Graduated from college of education? ( ) Yes ( ) No

Your level of graduation: ( ) Excellent ( ) Very good ( ) Good ( ) Satisfactory

No. of Years of Experience: ( ) 5 years or less ( ) 6 - 10 years ( ) Greater than 10 years

Please answer the following questions by circling your choice of (Yes) or (No) or by writing down appropriate numbers or texts in the specified spaces. For the multiple-choice questions, please circle the alternative of your choice.

Main Question	Answer	Sub Questions
1). Do you have a personal computer at home ?	**Yes**	Have you installed the CAIAT software on your own PC? **Yes No**
	No	Have you installed the CAIAT software on your school's PC? **Yes No**
2) If you have installed the CAIAT software on a PC, have you tried it out for discovering the quality of your tests' questions? (Regardless of whether you succeeded in this/these attempt(s) or not)	**Yes**	How many times? ……
	No	What are the reasons: …… …… …… ……
3) After attending the training course, has your use of HCD behavioural objectives for instruction increased?	**Yes**	How much do you estimate the ratio of increase to be? ……..%
	No	Would you please circle which of the followings was a reason for this (You may circle more than one and may write down other reasons that are not mentioned): a. Not interested in this issue. b. Do not have much time. c. This adds to my workload. d. The school does not appreciate such improvement in my ability.

		e. I feel afraid that I might make some scientific errors if I tackled HCD, thus I tend to be limited to the lower cognitive level. f. I think that teaching science/physics should not go as far as the level of higher cognitive demand. g. I did not understand how I can apply HCD concept implications in the real world. h. Although I understand what I have learnt on this course, I think I need more time until I understand it thoroughly and am able to apply it. Other reasons: …… …… …… ……
4) After attending the training course, have your questions at the level of HCD, either during instruction, in worksheets, in home work or in your tests, increased?	**Yes**	How much do you estimate the ratio of increase to be? …....%
	No	Would you please circle which of the followings was a reason for this (You may circle more than one and may write down other reasons that are not mentioned): a. Not interested in this issue. b. Do not have much time. c. This adds to my workload. d. The school does not appreciate such improvement in my ability. e. I feel afraid that I might make some scientific errors if I tackled HCD, thus I tend to be limited to the lower cognitive level. f. I think that teaching science/physics should not go as far as the level of higher cognitive demand. g. I did not understand how I can apply HCD concept implications in the real world. h. Although I understand what I have learnt on this course, I think I need more time until I understand it thoroughly and am able to apply it. Other reasons: …… …… …… ……
5) After attending the training course of HCD-CAIAT project, have you tried to use what you learnt about the CAIAT software on a self-learning basis?	**Yes**	How much do you estimate the ratio of your success to be? .....%
	No	Would you please circle which of the followings was a reason for this (You may circle more than one and may write down other reasons that are not mentioned) please put what is already in your mind: a. Not interested in this issue. b. Do not have much time. c. I am very confident that my ability in writing test questions does not need any improvement. d. The school does not appreciate such improvement in my ability. e. I did not understand what the project is all about. f. This adds to my workload. g. I did not feel confident that I can apply what I learnt in this course. h. I need some time until I understand thoroughly what I have learnt. i. I do not agree that this is a way to improve test item construction. More reasons: …… ……

Appendix 6

The Workshop Design

The workshop is held on two sessions. The first is a session of individual work in which every teacher uses the CAIAT software to analyse her/his experimental test. They also observed and evaluated by the research assistants. The second is a session of group work in which six groups of four individuals each is formed and asked to discuss their findings together. It is anticipated at this stage that some may have not been able to do proper interpretation of the analyses reports printed by the CAIAT, some may have found some difficulties in doing so, and some may need to be helped in how to get the analysis from the computer. However, such cases will uncover to what extent they have positively interacted with the project and the observation is anticipated to be rich and informative. Furthermore, they are going to gain more from interaction with peer teachers and potential ambiguity could be resolved via this interactive collaborative meeting. A kind of AR is aimed to take place within these groups, therefore I will encourage this by the following sheet of questions that may stimulate illuminating discussions and good reflection from the practice.

The Focus Groups Question Sheet

Dear Teacher,

This session is to help you communicate with your colleagues regarding the HCD/CAIAT project. In order to gain the most out of this meeting, you are invited to ask questions, say comments and share your experience with others. We have prepared some key questions that you may use for initiating discussions within your group. You are not limited to these and may ask other similar ones.

1. What are the main advantages of the CAIAT software?
2. What are the main shortcomings of the CAIAT software?
3. What are the key ideas of interpreting the CAIAT findings?
4. Are there any distinct examples of the way that the two indicators single out weak or good items?
5. Have any participants found any clue for better use of the project concepts (either for HCD or for CAIAT)?
6. What do the others say about their experience so far in making use of the project to increase their skills in constructing optimum HCD questions?
7. What do the others say about their experience so far in making use of the project to increase their skills in evaluating test questions by means of the CAIAT software?
8. What is the main value of focusing on HCD questions?
9. Do they think that the CAIAT software is easy to be used throughout the year?
10. Do they find that using the software instead of the manual calculations could encourage them to evaluate most of their tests?
11. Do they find that dealing with concepts of Difficulty Coefficient and Discrimination Index is easy to be handled?
12. Do they think that what these indicators reflect is reliable?

I will ask each group to discuss the issue on the basis of SWOT form (i.e. Strengths, Weaknesses, Opportunities and Threats). Their discussion should draw on their experience with the project rather than mere opinions and ideas. Thereafter, each group will present what they have come up with and a summary of the points that most groups have agreed on will be written on the board. The following table will be used for this task:

Group	Strengths	Weaknesses	Opportunities	Threats
1				
2				
3				
4				
5				
6				
7				
8				
Additional Comments: …… …… ……				

The additional comments may encourage the participants to clarify more points and issues concerning the application of HCD/CAIAT concepts.

Appendix 7

Observation Sheet During the Workshops

This research instrument is used by the research assistants during the first session of the workshop where every observer uses this form to observe 10 trainees or less while they are using the CAIAT software. The aim is to evaluate their skills on CAIAT by a sort of authentic assessment that based on a hands-on practice. The last five items are measured after the CAIAT session, when the teachers sit on a round table to comment on the results of their analysis and the research assistant share them this session.

---.

Dear Observer

Tick each selection according to your observation:

Teachers / Item of Observation	1		2		3		4		5		6		7		8		9		10	
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Do they have the ability to run the software?																				
Do they know the meanings of its menu components?																				
Do they know how to enter the initial specification of the test?																				
Do they know how to enter marks, using the software?																				
Do they print out reports easily?																				
Can they review entered information easily?																				
Can they edit any data entry errors easily?																				
Do they possess enough ability to single out the suspected weak test items upon their analyses results?																				
Do they possess enough ability to single out the good test items upon their analyses results?																				
Do they have enough ability to find out what is the probable cause of the suspected weakness for an item?																				
Do they have the ability to refine those weak items, using the criteria of optimum test questions?																				
Do they find any difficulty in comparing two coefficients to each other to gain one judgement? (For example, difficulty coefficient versus discrimination index).																				

- How many mistakes do they make usually? ……..

Please write down those mistakes:

……

……

- What main questions do they ask?

……

……

You may write down your other observations that the form has not included*:*

.....

.....

.....

.....

.....

Appendix 8

Summarizing Table for Observation

This research instrument is used by each research assistants to summarise what they had found during the workshop for the set of teachers that they were asked to observe. This finally is also summarised to give an overall finding.

.----------.

Tick each selection according to the averages of your observation:

Observational guides	Yes	No	Not applicable	If no, how many persons cannot?
Do they have the ability to run the software?				
Do they know the meanings of its menu components?				
Do they know how to enter the initial specification of the test?				
Do they know how to enter marks, using the software?				
Do they print out reports easily?				
Can they review entered information easily?				
Can they edit any data entry errors easily?				
Do they possess enough ability to single out the suspected weak test items upon their analyses results?				
Do they possess enough ability to single out the good test items upon their analyses results?				
Do they have enough ability to find out what is the probable cause of the suspected weakness for an item?				
Do they have the ability to refine those weak items, using the criteria of optimum test questions?				
Do they find any difficulty in comparing two coefficients to each other to gain one judgement? (For example, difficulty coefficient versus discrimination index).				

- How many mistakes do they make usually? ……..
Please write down those mistakes:

.....
.....
.....

- What main questions do they ask?

.....
.....
.....

You may write down your other observations that the form has not included*:*
.....
.....
.....

Appendix 9

Some Research Instruments

A checklist for lesson observation

#	Items	Yes	No
1	Do they use behavioural objective from HCD levels?		
2	Are these behavioural objective stated well?		
3	Do they ask HCD questions during instruction?		
4	Are their HCD questions of good quality?		
5	Are these questions eliciting some challenge amongst pupils?		
6	Do they follow good techniques in asking such questions?		
7	Do they follow good techniques in receiving pupils' responses?		
8	Do they follow good technique in reinforcing pupils' responses?		
9	Are their pedagogical practices suitable for teaching thinking requirements?		
10	Are the pupils interacting actively with questions of this type?		
11	Are pupils able to give answers?		

A checklist for content analysis of teachers' tests

#	Items	Yes	No
1	Are the test questions of HCD levels?		
2	What is the ratio of those that relate to HCD?		
3	Are there any defects in stating the questions?		
4	If there are defects, are they from the type that IAT could indicate to?		
5	Are there answers that indicate quality of learning?		
6	What is the percentage of this type of answer out of all?		

A Guide for open response interview

1- What is your overall perception about the project's main components: HCD and CAIAT?
2- Have you got any benefit?
3- What is the main value of focusing on HCD questions? For example what are your observations of effects of the HCD dimension revealed by the project on your pupils' learning?
4- What difficulties you have faced during the project?
5- Do you find that using the CAIAT software instead of the manual calculations could encourage you to evaluate most of your tests? If "no" why?
6- Do you find that dealing with concepts of Difficulty Coefficient and Discrimination Index is easy to be handled?
7- Do they think that what these indicators reflect is reliable?
8- What are your suggestions?

Appendix 10

Case Study Instruments and Findings

Instruments

The case studies' aim is to discover the extent to which teachers were able to apply concepts and skills of the project. It includes three research instruments: (1) content analysis of the teacher's paper work, (2) observation of how do she use the CAIAT, and (3) interview about her analyses of the findings and about the whole experience in general. The following questions were held in mind to be answered for each of the three instruments:

Content analysis

1. Are the teacher's test questions of HCD level?

Observation

1. How easily can she install the CAIAT software on a PC?
2. How easily can she use the CAIAT software?
3. If she to face a problem during her work, how she would solve it?
4. To what extent she was quick in entering the data?
5. What mistakes she could make during her work?
6. Any other?
7. Can she discover weak or optimal items in terms of IAT concepts?
8. Would she initially describe reasons for the strength or weakness of the items?
9. What acts would the teacher undertake, or suggest, for finding out reasons for contradictions that she might find when reading the CAIAT software item analysis results? (Examples of these might be: reading through pupils' answer-sheets or meeting some of the pupils).

Interview

1. Can she elaborate on reasons for strength or weakness of items?
2. Would she suggest modifications to any weak items where necessary? What would these suggestions be? (if she did not suggest initially, the observer will ask for this)

3. What lessons/benefits can she utilise as a result of her new practice? In this term, to what extent can she recommend employing item analysis results for improving her teaching skills or ability for test construction?
4. What are her comments? (There could be other subsequent questions that emerge from comments)

Findings

The presentation of a case starts with an introduction to the case which presents the case teacher, the case period of time and the place, the school atmosphere – especially the attitude of the principal and her assistant – and the atmosphere of the CS situation. Next, an illustration of the teachers' use of the CAIAT software shows how she uses the software, by addressing the main observations that clarify her level of ability. Finally, an illustration of IAT being applied by the teacher will be composed of three sections. First, an evaluation of the teacher's test questions, in which I provide grounds for the reader by describing her test in terms of its number of questions, types of questions and HCD analysis. Second, the analysis results, in which I explain briefly the test item analysis results. Third, the teacher's reading of these results, which presents the interpretations, comments or suggestions the teacher has produced.

To clarify what is mentioned in the third section above, this relates to what the teacher has highlighted, rather than going over all the questions and asking her to evaluate or comment on each one. This allows the teachers to select on their own what they believe to be weak or good items and then to receive their comments that fit with what the CS seeks to explore, where it reveals the priorities that each teacher has and what her ability is for noticing those items that needs to be improved. I then discuss what has been selected in terms of the extent to which the comments are relevant to IAT concepts and the criteria/considerations of HCD question construction. Appendix Table 10.1 illustrates the major findings in a summarised form.

Appendix Table 10.1: Summary of the case study findings

Item	1st case	2nd case	3rd case	4th case	5th case
Key stage	Intermediate	Intermediate	Secondary	Intermediate	Secondary
Subject of specialisation	Chemistry	Physics	Physics	Chemistry	Biology
Years of experience	13 years	6 years	6 years	13 years	6 years
Professional development through prior relevant training	Modern Pedagogy	1- Computer diploma 2- Construction of achievement test	None	None	1- Construction of achievement test. 2- Intermediate science pedagogy
Teaching level	Very high	High	Moderate	Moderate	High
Principal interest	Low	Very low	High	Very low	Very High
Atmosphere of the school	Quite normal: no critical follow-up, or opposition from the principal or the assistant principal.	- Some tensions. - The school administration did not appear to be interested in the educational issues or in this project. - Low pupils' achievement results.	Friendly: excellent relationship between the principal and the teachers. - The principal welcomed the case study. - Careful about pedagogy. - Cooperative approach to the project's application.	NA	- The principal is supportive. - Welcomed the application of the project with her teachers. -excellent relationship among the principal and her teachers.
Atmosphere of the interview	Welcoming: was talking enthusiastically about the software and showed her pleasure with this during the case study.	Welcoming: showed her happiness at being selected for the study and cooperated very well.	- Was a bit confused about the situation. - This affected the level she performed at. - Later this apparent nervousness disappeared.	The teacher's interest in professional development is limited although she has a good level of instruction.	The teacher is happy with her experience. Active, efficient and keen to try out new things. Has a keen interest in the project. Responded very well to the project investigating activities.
Time needed for the case study	3:30 hours	4 hours	3 hours	4 hours	3 hours
Venue of the case study	The school library	The school library and the assistant researcher's office	The principal's office	The school library and the assistant researcher's office	The principal's office
Findings Summary for Effectiveness Dimension					
Level of running CAIAT	*Very High*	*Low to Moderate*	*Moderate*	*High*	*Not available*
Level of IAT interpretations	*High*	*High*	*Very High*	*Low to Moderate*	*Low to Moderate*

First Case

Introduction to the case

This is a chemistry teacher who has 13 years of experience in intermediate school and holds a Bachelor's degree in Chemistry. She has been trained in teaching methodology for intermediate school science and a course in modern pedagogy. The time required for the CS was three and a half hours. It was conducted in the school library after the teacher had finished marking a chemistry test at the end of the second semester.

This teacher is one of the outstanding teachers at the level of the province's schools, according to her principal and the educational supervisor. Moreover, according to the research assistant who conducted the CS, she is a serious and hard worker. This was seen when she faced an obstacle when trying to run the software: when she attempted to use the software she was unable to run it on the first computer but continued to try it on other computers until it worked. She also asked for assistance from her husband who specialises in a related subject. She expended significant efforts to transfer her experience to her peer teachers after returning from the training course for this project. She has a quiet, confident and wise personality.

The school atmosphere is a normal one, in that there is no critical follow-up or support from the principal or the assistant principal and neither was there any opposition. They were not keen to look into the project's details or its purpose. For research purposes this is not a shortcoming but rather a positive circumstance, because the presence of the principal's interest and/or support might confuse the teacher's self-impetus and could thus colour the findings of the study. The interview atmosphere was a welcoming one, where the teacher talked enthusiastically about the software and its purpose. She was forthcoming with her suggestions to improve it and was focused, precise and showed her pleasure with this during the CS.

Findings

Use of the CAIAT Software

This teacher was overwhelmingly appreciative of the importance of the CAIAT package and its role. She showed an obvious interest in the details of the package. During her use of the CAIAT, she was quick and accurate in following the steps of entering the initial data such as adding a class, subject or pupils' names. She raised some new issues in using the software, such as copying and pasting data from one location to

another. Using trial and error, she identified where exactly one can perform such operations. She indicated that it could be done during item name entry only, while other data entry functions do not allow such an operation. This indicates her mastery of using the package, which most probably, occurred as a result of her intensive use after the training period. All of these aspects are indicators of what Q6d seek to explore.

To enter the marks for the items quickly and effectively, she printed out a specially designed table which included pupils' names and the corresponding marks for each item. Moreover, she used the 'repeated mark' data entry function of the software, in which the common mark is entered by repeating it with this function, and then went back over the records to change the few marks that were different.

One of the strongest pieces of evidence of the high level of interest that this teacher had for the use of the package (and very likely about its underlying concepts and implications) was her list of suggestions to improve the software. Her suggestions were:

> - In the data entry portion of the software for each item, it is suggested that the number of the item appear besides its description.
> - When entering the initial data the user should be aware of the accuracy of the data so as not to cause an error in all subsequent operations because they were based on those initial data. In such a situation, the suggestion is that the program should discover this sort of error and alert the user before giving out its false results.

Consequently, I would classify her as an optimal user of the software, who has interacted with its purpose actively enough to give a good reliable judgement and reflection about its usage and functionality both from her verbal responses and her observed behaviour. For Q6d, she acquired a significant level of skills for optimal use of the CAIAT software.

Applying IAT Skills

a. Evaluation of the Teacher's Test Questions

The test was composed of 30 items. Some were subjective and others objective (of three types: multiple-choice, matching and short answer.) Of the 30 questions, there were 10 at HCD level with a ratio of 33%. Levels of cognition were comprehension, application and analysis. For example, there was a question asking the pupil to use reason to explain some phenomena, another was to compare the Carbon cycle and Nitrogen cycle, and a third question asked the pupil to analyse the characteristics of some plants in order to identify their type.

b. Analysis Results

Difficulty coefficients are between 53% and 99% with an overall average of 83% which indicates an easy test generally. 27% of the items are of the good difficulty coefficients between 50% and 75% and the rest are very easy ones above 75%. Discrimination index values fall between 2% and 69% and generally they are not high as only 27% of these are of good discrimination above 30%.

c. The Teacher's Interpretations

The teacher was able to read the results of the CAIAT software outputs accurately and succeeded in correlating difficulty and discrimination parameters. Item 1 asks the pupil to give the scientific term that matches the expression "Sorting live creatures regularly according to their food production and consumption" and received a difficulty[50] coefficient of 98% and discrimination of 4%. The teacher reasoned that this decline in parameters was because of the easiness of the required skill for the answer, which was recall. She adds that the extreme easiness of this item comes from the lack of good construction of the question, where there was a wording relationship between the scientific term that the correct answer should include and the expression that is shown in the text of the question. This facilitated a path for the pupils to select the correct alternative without necessarily possessing the knowledge of that content. Similarly, items 2, 3 and 4 were of a high difficulty and low discrimination and she reasoned this by the fact that these items required only simple recall, as they included scientific terms for which the definition for each is unique and there are no similarities that could aid the efficiency of distractors. The result of this reading is that she mentioned that her next tests should not ask about scientific terms in multiple-choice type questions. She thinks that the alternatives to this are:

> - It is better not to mention the scientific terms but mention their descriptions and ask the pupil to write down the term, i.e. to transfer from multiple-choice to short answer question.
> - Alternatively, one could select all scientific terms of all items from one chapter where there could be some similarity among them and this aids having good functioning distractors and avoids wording relationships between the correct answer and the text of the question.

Item 13 is a multiple-choice one that asks the pupil to identify from which two materials the word SIMA comes from, out of three alternatives: Silicon and Nickel, Silicon and Magnesium, and Silicon and Aluminium. The discrimination index for this

item was 2% and she reasoned this by the fact that the wording pitfall in this item is very obvious and has affected its results. She decided that this question should not be a multiple-choice type next time.

On the other hand, the teacher succeeded in identifying the good items that gained optimal parameters. For example, she denoted item 15 as a good item. It asks the pupil to identify to which group of plants the colocynth plant belongs. The teacher considered that because the skill required for answering this item is within the level of analysis, she decided to keep it as a good item since its two parameters were high where the difficulty is 53% and the discrimination is 44%. Similarly, she classified item 18 as an analytical question for comparing Nitrogen and Carbon cycles and decided that it was an optimal question since its two parameters were in the good range of values at 66% for difficulty and 35% for discrimination.

Item 16 says that: earth completes its round around the sun each: (24 hours, 29 days or 365.25 days). The two parameters were 82% for difficulty and 27% for discrimination. It was anticipated that the difficulty would be more than this value because the question content was far too easy from the perspective of the teacher. However, she justified this result by pointing out that the phrasing of the item's text was less efficient because the term 'its round' might have confused a number of pupils. She strengthened this inference by the finding she received about the distractors since most of the pupils who failed this question selected alternative b (29 days). I asked the assistant researcher to clarify this point to me by asking the teacher why she thought this selection was confusing. She did so, and received the response that the word 'day' was the reason. Because of the word 'round', the most popular round is when the earth goes around itself every 'day.' However, I think this is not the only possible reason, as the third alternative has the word 'day' as well. It is more likely that the choice of the number 29 has another possible and parallel reason, because it is also commonly known that the moon completes its 'round' around the earth every month, for which 29 days is a good representation. The popularity of this fact in the Saudi context does not come from the curriculum content in the first place, but is part of Muslims' worship, in relation to the month of Ramadan, the month of fasting, and month 12 for going on pilgrimage to Mecca. The starting days for both occasions are identified according to the movement of the moon. In cases like these, I do not go back to the teacher and raise these issues to receive a new justification or interpretation from her. There are two main reasons for this: firstly, so there will be no further 'questioning' which might make her

feel threatened, and secondly, because it is the first opinion that the research subject gives that is the targeted response, not the ‘defensive’ or ‘justified’ later interpretations that could have benefited from new questions or comments. However, the teacher's interpretation is still valid as one possible reason, which is a good start for a beginner; especially that she connected her reasoning to the fact that most of the pupils who failed this question selected alternative b that includes the term 'day'.

Item 30 has a 64% for difficulty and a 69% for discrimination and hence she classified it as a good item that should be reused latter. Similarly, she decided to keep items 7 and 24 as good items because their parameters were in the good range of 70% and 60% for the first and 74% and 27% for the second.

My conclusion from the preceding presentation of this teacher's interpretations is that she highlighted successfully a number of items with a very relevant use of IAT, showing a good level of understanding of what the project's training aims are in terms of IAT skills. This should answer the research question Q6e positively. Furthermore, her justification about the reasons that underlie some weak questions reflects the good level which she has acquired for the skills of writing HCD questions which are targeted by research question Q4c. She commented using different conceptual frameworks: Bloom taxonomy, wording errors in constructing multiple-choice items and distractor construction.

Second Case

Introduction to the case

This is a science teacher in an intermediate school with 6 years of experience, who holds a Bachelor's degree in Physics. She received a diploma on the use of computers and training for the construction of achievement tests. The total time needed for completing the CS was four hours. The first part of it was undertaken in the school library; this was devoted to a discussion concerning the test questions and to observe her work with the software. The second was after marking the test at the end of the second semester, and was conducted at the assistant researcher's office.

She paid careful attention to the issue of professional development, seeking new methods and looking forward to trying out new things. She prepared good worksheets and kept using the teaching aids in her lessons. She was enthusiastic about the project and was very interested in knowing much more about CASE; she had even visited the CASE website after she had heard about the concept during the project training session.

The school atmosphere has some tensions and the school administration did not appear to be interested in the educational issues or in this project's activities. Indeed, the principal did not ask about the purpose of the project team's visit. The pupils' achievement results in this school are generally low level. However, the interview atmosphere was welcoming and the teacher showed her happiness at being selected for the study and cooperated very well with the assistant researcher.

Findings

Use of the CAIAT Software

This teacher's use of the software was generally moderate. Although she is not a professional user and had faced some minor problems, she was able to continue and enter data. The program menu and the meaning of its contents were clear to her, and her speed in entering data was moderate with few errors. She was able to open a class file but needed more than one attempt to succeed; she then concluded by adding the initial data of the pupils and items successfully. However, there were some times when she forgot to enter the item standard mark or to define its type. This is because she did not attempt to use the software more than once during the contingent stage of the project, which was after the training session. Finally, although she was not classified as a professional user of this software, she succeeded in obtaining a report of the item analysis and was able to make use of the software on her own.

For research question Q6d, I consider that her level for using the CAIAT is low to moderate, which reflects a satisfactory acquisition of the project aims in terms of running the CAIAT software successfully.

Applying IAT Skills

a. Evaluation of the Teacher's Test Questions

The test was composed of 24 items. Some were subjective and others were objective (of four types: multiple-choice, true-false, completion and short answer). Out of the 24 questions, there were 6 at the HCD level, representing a ratio of 25%. The levels were comprehension and analysis. For example, there was one question asking the pupil to reason and another asking him/her to compare. Obviously most of the test items required recall, which was a matter of discussion with the teacher where she agreed that her future tests should include more HCD questions. However, she raised the issue of self-training beforehand, so as to be able to write such questions, thus

highlighting the role of the CAIAT software in this regard. She also mentioned that asking thinking questions during instruction is a requirement for eliciting learning.

b. Analysis Results

Difficulty coefficients were between 61% and 100% with an overall average of 82%, which indicates an easy test generally. 42% of the items were of a difficulty coefficient between 60% and 75% which is considered to be in the moderate level of difficulty. The rest were very easy items, above 75%. Discrimination index values were between 0% and 66% and generally they were moderate since 50% of the items were of good discrimination, above 30%.

c. The Teacher's Interpretations

She was able to discover the good items such as 10 and 12. For item 10, which asks for the characteristics of a mountain environment, the level of difficulty was 70% and discrimination was 53%. She justified this good result because of the nature of this item, which tends to be analytical. For item 12, which requires pupils to mention two components of the life environment, difficulty was 61% and discrimination was 66%. However, she states that this item is actually easier and hence the D value should have been more than 61%. Her reasoning for this is that during the instruction, the pupils were required to mention only one component of the life environment, which is living creatures, while this question asks for two components. Her feedback from this was that she might use thinking maps to present the fact that there is more than one life environment component.

She classified items 15 and 17 as good items because they had 69% and 72% respectively for difficulty and 38% and 38% respectively for discrimination. She indicated that this was due to their level of cognition being comprehension reasoning questions. Item 5 was a good question of 81% difficulty and 38% discrimination. She thought that a good indicator that supports the statistical parameter quality judgement is that the pupils gave a new answer that was not expected. The question asks the pupil to identify whether the text is true or false and then to correct the false answer. The text for this item was “Sandy soil is rich with salts and argil.” Some pupils, instead of mentioning what materials the sandy soil includes, mentioned that the correction is “Mud soil is rich with salts and argil” which is correct as well.

She also denoted some weak items in terms of the parameters' indications. Item 22 has a 100% difficulty and 0% discrimination, which represent a very easy question. She suggested that this item's distractors should be changed so that it works better. Both items 2 and 8 have similar values: a 97% difficulty and 6% discrimination. This was a result of her presentation of the related lesson for each of the two questions in which she used a scientific film that explained the subject. She thought that this added distinctively to the pupils' comprehension as well as their memorization of the topic's content and thus was reflected in their answers this way. She thought that the benefits she achieved from item analysis of her test were:

1. Using scientific films has a distinctive role in aiding pupils' comprehension and recall of information.
2. It is worthwhile asking pupils to undertake some research work about some topics of the curriculum.
3. Varying question types is an important issue during test construction, especially for those at the application level.

My conclusion from the preceding presentation of this teacher's interpretations is that she highlighted successfully a number of items with very relevant use of IAT and has shown a good level of understanding what the project's training aims are in terms of IAT skills. She also connected what she found to her instruction practice, which should provide her with good feedback and on-going improvement in terms of professional development. This should give the answer to research question Q6e. Furthermore, her justification about the reasons that underlie some weak questions reflects the extent to which she has acquired good level of skills for writing HCD questions, which is targeted by research question Q4c; in this respect, she commented using two important conceptual frameworks: Bloom taxonomy and distractor construction.

Third Case

Introduction to the case

This teacher is a Physics secondary school teacher with 6 years' experience, who holds Bachelor's degree in Physics from a College of Education. Observation and interview were conducted in the principal's office where the school's PC is located. This interview took place for three hours after the end of the second semester examinations and after the Physics test papers had been marked. The level of the test questions' quality and the data analysis results had been discussed with the teacher. The school principal pointed out the extra time and effort that the teacher spends, including her

active contributions towards extra-curricular activities, and her keenness to apply what she has learnt from this project's training. However, during the interview she showed a low ability generally. The school's atmosphere is characterised as being friendly and the relationship between the principal and the teachers is excellent. The principal welcomed the CS and indicated an interest in its implications and the skills included. She is careful about issues related to instruction and cooperative in terms of the project's application requirements.

During the interview the teacher was somewhat confused (or maybe frightened about the situation), which affected the level she performed at while running the software. She needed many attempts before she was able to initialize a new class and subject. However, later this apparent nervousness disappeared and she was able to perform at a higher level.

Findings

Use of the CAIAT Software

This teacher's use of the program was generally at a moderate level. She is not a professional user and she faced some problems using the software. However, she was able to continue and enter data. At first, she was not able to initialise a new class, and only succeeded after a number of attempts. She entered a set of pupils and the initial data items but her work was not very accurate: she missed entering the item standard mark during the "add new item" function. However, she discovered her problem and was able to modify the items by means of the "edit item data" function. Moreover, during data entry, she forgot to enter an item mark; but later she noticed that the sum of all the marks was not the same on that pupil's paper. Consequently, she discovered the missed mark and entered it. Her overall data entry speed was moderate, with the exception of the multiple-choice items, which was fast due to the similarity of alternatives that has been chosen by most of the pupils. In general, I consider this case as an exemplar of the beginning point for most teachers, especially during their first attempts at using the software. For research question Q6d, I consider that her level of using the CAIAT is moderate, which reflects good acquisition of what the project aims are in terms of running the CAIAT software successfully.

Applying IAT Skills

a. Evaluation of the Teacher's Test Questions

The test is composed of 34 items. Some were subjective and others objective (of two types: multiple-choice of three alternatives and a short answer). Out of the 34 questions, there were 10 at HCD level at a ratio of 29%. The levels were comprehension and analysis. For example, there was one question asking the pupil to reason and another asking him/her to compare, as well as a third one asking the pupil to explain.

b. Analysis Results.

Difficulty coefficients were between 56% and 100%, with an overall average of 90%, which indicates a very easy test generally. 18% of the items had a difficulty coefficient of between 63% and 75%, which is considered to be a moderate level of difficulty. Almost 80% of the items were easy ones, with 16 items of 100% difficulty value, which represents almost half of the test (47%). Discrimination index values were between 0% and 50% and generally they were poor, since only 7% of the items were of good discrimination, above 30%. All the 100% difficulty value items have a zero discrimination value, which leaves the test with almost half of its items (47%) as non-discriminating items.

c. The Teacher's Interpretations

She was able to read the analysis report and correlate between the two parameter values. She commented on the decrease of some of the item discrimination statistically as a result of their high difficulty value, or being very easy ones. She noticed an item of a negative discrimination value but reviewed the extent to which the data entry process was correct and found some errors; she then modified these, re-analysed the data and hence obtained a positive value. From reading the two parameter values, she was able to select the good items where she classified item 31 in this category because it has a 69% difficulty and 63% discrimination, then she read the item distractor indicators where she found that all were good. Similarly, she classified items 8, 14 and 24 as good items since their difficulty and discrimination were 75% and 50%, 63% and 25%, and 56% and 38% respectively. I read through these items to evaluate their difficulty and then judged her reading of the indicators. My appraisal is that the first two items are not easy, since both require reasoning and their content is not a very common idea which pupils usually keep in focus. The third item is a problem-solving question, which can be difficult. I have to point out that since the indicators show a high rate of correct answers it indicates good teaching practice. She highlighted them as being good items,

which is true within this sense. On the other hand, she also correlated between the two indicators, difficulty and discrimination, justifying her judgement on the items' strength.

She criticised item 15 albeit its parameters were 63% and 50% because she anticipated that the content of this item should be very easy for pupils and the difficulty should have been at a value of more than 63%. However, she justified this result by the fact that the content comes in two different locations within the textbook, which might have distracted the lower-achieving pupils. In my opinion, this justification is reasonable and can be true in some sense. She reflected that pupils should be alerted to this issue in the feedback session of the test.

She commented on the items that were of 100% difficulty and 0% discrimination as being weak items because most of them were at recall level. However, there were some problem-solving questions, but she indicated that they had been explained many times to the pupils, which could have made the pupils' cognition of this curricular content similar to that of recall level. Some other items were multiple-choice (26, 27, 28, 29, 30 and 34) on which the teacher commented that their alternatives should be revised in order to make them work more efficiently. Item 20 was of distinguished discrimination of 0% whereas difficulty was 75%, which was noticeable and revealing a paradox. She went back to the pupils' answer-sheets so as to see where this result came from and found that the pupils had written another name for the instrument they were being asked about. She also gave attention to the distractor efficiency for multiple-choice questions. She found that the efficiency coefficient of one alternative for item 25 was of 0% value. This ratio is calculated by subtracting two values of the higher group selectors for this alternative and the lower group selectors from each other. These two groups' values for item 25 were equal; hence the zero value of the distractor efficiency did not come from two zero values; which indicates that this alternative could be true in some sense because both groups of higher achieving and lower achieving pupils have chosen it equally.

My conclusion from the preceding presentation of this teacher's interpretations is that she highlighted successfully a number of items with very relevant language for IAT and has shown a high level of understanding what the project's training aims are in terms of IAT skills, especially as she appreciated the practice of reviewing the pupils' answer-sheets, which reflects the rich experience of the teacher and highlights reflection on some of her teaching practices which also counts for the professional development reflection aspect of the IAT exercise. Moreover, she commented successfully on the 0%

value of item 25's one distractor efficiency which reflects a good understanding of the statistical vision of this subject. All this should answer research question Q6e, that her acquisition of IAT skills was very high. Furthermore, her justification about her reasons that underlie the weakness of some questions reflects the extent to which she has acquired the skills of writing HCD questions, which is targeted by research question Q4c. In this term she connected her opinion to the curriculum for item 15, to the distractors' efficiency for item 20, and to Bloom taxonomy for some other items. Nitko (1983) indicated that IAT could aid teachers in understanding problems coming from curriculum.

Fourth Case

Introduction to the case

This is a science teacher in an intermediate school who has 13 years of experience and holds a Bachelors' degree in Chemistry from a College of Education. Observation and interviews were carried out at the school library and at the assistant researcher's office. This took place for four hours in two sessions after the end of the second semester examinations and after the science test papers had been marked. The level of the test questions' quality and the data analysis results was discussed with the teacher. The teacher's interest in professional development is limited although she has a good level of instruction. She believes that the low level of the pupils' achievement is the reason for not being able to increase the cognitive level of the questions or to target their thinking skills during instruction.

The school atmosphere is not stable, with low pupil results. The principal did not ask about the purpose of the project team's visit, which revealed that less attention was given by the school management. However, the interview atmosphere was good where the teacher was happy to join the study and thus cooperated very well with the assistant researcher.

Findings

Use of the CAIAT Software

This teacher's use of the program was generally good. She is not a professional user but she succeeded in initializing a new class and adding a set of pupils and the initial data set of items. The meanings of the menu items of the software were clear to her. However, she faced a problem during the setup where she had to try to use the

software on more than one PC until she was able to succeed. I have to mention that the CAIAT setup has this problem with some machines, depending on the version of Windows or what other packages are installed on the machine, thus this problem was not usual for many participants and was not related to this particular case but rather was a common difficulty throughout the project. She did not try out the software more than once during the contingent stage of the project which was after the training session. During the entry of the items' marks she was slow though she used a specially designed form that has the marks on a table to aid her in entering the data very quickly. In the end she was able to analyse and obtain the report of the analysis results successfully.

For research question Q6d, I consider that her level of using the CAIAT is high because although she did not pay close attention to the software during the contingent stage, she showed a good level of ability for running it, which could be a result of her achievement from training. This reflects the extent to which the project training was successful in fulfilling its objectives in terms of running the CAIAT software and also highlights the ease of use of the CAIAT software.

Applying IAT Skills

a. Evaluation of the Teacher's Test Questions

The test was composed of 22 items. Some were subjective and others objective (of these types: limited multiple-choice of two alternatives, completion, matching and short answer). Out of the 22 questions, there were 18 recall type questions of a ratio of 82% and the rest were of HCD level, which represents only 18% of the test. These HCD levels were application and analysis. For example, there was a question asking the pupil to reason and another asking him/her to compare as well as a third one asking the pupil to classify.

The teacher indicated the need for her test questions to include further HCD items in the future. However, she believed at the same time that this should be preceded by making the pupils accustomed to questions of this type. She is planning to work on this objective next year.

b. Analysis Results

Difficulty coefficients were between 54% and 100% for an overall average of 76%, which indicates an easy test generally. 46% of the items were of a difficulty coefficient between 50% and 75%, which is considered to be at a moderate level of difficulty. The

rest were easy items of difficulty more than 75%. Discrimination index values were between 0% and 88% and generally they were good since 73% were above 30%.

c. The Teacher's Interpretations

Her technical reading of the analysis results was not high because she classified the items according to their difficulty coefficient without looking at the discrimination. When the assistant researcher asked her why, she replied that she did not understand what the boundaries or criteria were for the quality of the discrimination parameter. Nevertheless, her explanations of the results are generally satisfactory since she was able to justify her opinions by the standards for question construction; thus her ability in terms of IAT tends to be moderate.

The difficulty of item 6 is 54%, which she considered a reasonable value since its content requires the pupil to analyse so as to discover the corresponding classification of onion root types. Similarly, item 12 received a 60% difficulty value which she interpreted by the fact that a plant cell contains various elements, thus each could probably confuse pupils when attempting to identify the corresponding function for one of them according to what the question requires. Item 19 has a 60% difficulty and 52% discrimination. She commented that this value of difficulty is anticipated for this item since it includes classification which requires HCD level skills.

The difficulty value of item 13 is 71%, which she considered above her anticipated value because the skills and knowledge required for the answer are of a higher cognitive order and require many steps in order to be solved. However, she justified this high difficulty value (i.e. being easy question) by the way she presented this subject during her lessons where she depended on a story-telling technique as she presented Archimedes' story. This could have strengthened pupils' comprehension of the underlying concepts and their retention of related facts and ideas. I think that this is one of the most important aspects of professional reflection that teachers can gain from this semi-action research practice. Adversely, she denoted that item 19's difficulty value of 56% is not high enough for what she was anticipating according to the overall easy level of the content. She suspects that the linguistic difficulty of the term ‘taxis’ is a probable cause of this result. I think this is not necessarily the reason because the terminology should be part of the skills or knowledge targeted by the curriculum and pupils are very likely to be aware of this relationship during their learning or preparation for testing. However, the teacher’s expectations might have been too optimistic

compared to her low quality instruction for this part of the curriculum. I think that as her future trials and analysis focus on the reason she suspects, she will find that she still has the same issue. She is most likely at that time to change her mind about this justification and look from another angle which relates to her instructional practice rather than from this specific piece of reason which she outlined. The educational supervisor's role in this sort of practice should help in promoting better scaffolding for the teachers' learning.

The teacher was able to identify weak items according to the analysis results. She identified items 20 and 21 as weak items according to their 100% and 91% difficulty respectively and interpreted this weakness as being the result of using multiple-choice questions type of two alternatives for fact recall skill questions, thus she thinks that in future they should be written in short answer form in which the pupil will write down what she is supposed to be recalling. She also suggested changing the root of the question or to increase and improve alternatives. For item 9, she thinks it should have had a difficulty less than the 91% it has. The reason is that since the item says “Try to discover where the wrong idea in this statement is and then correct it: The spring is an elastic body, therefore it is affected by a permanent deformation” she thought that the pupils might need to identify first to which type of material classification a spring belongs before they decided how to answer, but what happened was that the error in the statement was obvious and could be discovered easily without following the classification path that she was thinking about. Her suggestion for improving the item was to change its type from correction to multiple-choice or completion, which I think is reasonable. She thought that the value 88% of item 7 difficulty was too high and hence the item was less effective as a very easy item. In this regard, she indicated the easiness of the question task, which was matching the sections of a diagram to their key terms. Her suggestion for future improvement was to make this question a completion type in which the drawing is to be shown and the pupil is required to write down the corresponding key terms in the corresponding space. She also suggested changing the question into a full subjective type in which the pupil is required to draw the cell.

She was keen to look through pupils' answer-sheets for further understanding and applied this to item 14 which was of 69% difficulty, where she found that one important reason for this value is that the pupils who did not answer correctly wrote the benefits of classification instead of the basics of classification as the question asked.

Therefore, she thought that this item had a proper difficulty value which fitted with the pupils' level of achievement and abilities.

The teacher ended up with these lessons:

- She has to use further experiments and show scientific films to aid retention of information and facilitate comprehension.
- Her questions should be from different types of questions with a concentration on HCD and she should appreciate that each type of question fits much better with specific types of skill or knowledge.

My conclusion from the preceding presentation of this teacher's interpretations is that she highlighted successfully a number of items with the relevant language from IAT and showed a good level of understanding what the project's training aims are in terms of IAT skills, especially as she appreciated the practice of reviewing pupils' answer-sheets and highlighted the reflections of some of her teaching practices, which counts for the professional development reflection aspect of the IAT exercise. This should answer research question Q6e. Furthermore, her justification about the reasons that underlie the weakness of some of the questions does not reflect a good level of acquisition of HCD question construction skills, which are targeted by research question Q4c. This is because she focused on connecting between the levels of cognition and the types of questions as a framework for judging the items' weakness, and on improvement suggestions, which are the easiest and most fundamental approach for improving HCD questions. Looking at the language of the question is more important in this regard and reflects better ability and experience.

Fifth Case

Introduction to the case

This is a biology teacher in a secondary school who has 6 years of experience and holds a Bachelor's degree in Biology from a College of Education. She has been trained on achievement test construction and attended a training course on intermediate science pedagogy. The total time for the CS was three hours and was undertaken in the principal's office after marking the test at the end of the second semester. It was devoted to observations of her work on the software and a discussion about the test questions and results analysis. According to the principal, this teacher is active, efficient and keen to try out new things. She has a keen interest in the project to the extent that she, on her own initiative, invited her educational supervisor and the assistant researcher to attend

one of her lessons. She responded very well to the project investigating activities and was happy with her experience. The principal is a supportive person who welcomed the application of the project with her teachers and the CS being done with them. The relationship between the principal and her teachers is a kind one. She commended the teacher's interest in applying the project's concepts and skills during her work and followed up what was going on.

Findings

Use of the CAIAT Software

Unfortunately, this teacher's use of the CAIAT software has not been reported because the teacher carried out an item analysis by the CAIAT software in the absence of the research assistant. This is because she thought that the CS was concerned with the IAT sessions only.

Applying IAT Skills

a. Evaluation of the Teacher's Test Questions

The test was an experimental test and not an official one. It was composed of 8 items. Some were subjective and others objective of short answer type only. All of the 8 questions were of the HCD level. The HCD levels were analysis, synthesis and evaluation. For example, there was a question asking pupils to compare and another asking them to draw a thinking map.

b. Analysis Results.

Difficulty coefficients were between 64% and 80% with an overall average of 74%, which indicates an easy test generally. 37.5% of the items were of a difficulty coefficient between 64% and 70%, which is considered to be a moderate level of difficulty. The rest were easy items of a difficulty level at more than 70%. Discrimination index values were between 17% and 57% and generally they were good, since 63.5% were above 30%.

c. The Teacher's Interpretations

She was able to read the results of the CAIAT software outputs accurately and succeeded in correlating difficulty and discrimination parameters. She also identified good items upon their parameters. For example she identified item 1 as a good item because its parameters are 86% and 38% respectively. She commented that since this

item is about asking to draw up a thinking map it tends to be difficult and should have received a value of difficulty coefficient less than what resulted. However, she justified the higher result it had (i.e. that it is being an easy item) by the frequent training she delivered to her pupils for this skill. Likewise, she classified items 3 and 4 as good items upon the two values of 71% and 57% difficulty and discrimination respectively for both of them. She again commented on the increase in the difficulty value although the content of this item was not easy, as the illustrative drawing that she used as a teaching aid contributed substantially to the pupils' understanding of the related concepts and ideas, and this aided in the long lasting retention of this information, which was then reflected in the answers.

My conclusion from the preceding presentation of this teacher's interpretations is that she highlighted a number of items satisfactorily with the relevant language from IAT and has shown a moderate level of understanding of what the project's training aims are in terms of IAT skills. This should answer research question Q6e. Her justification about the reasons that underlie the weakness of some of the questions did not reflect good acquired skills for the writing of HCD questions which was targeted by research question Q4c.

Appendix 11

Raw Data

The Questionnaire raw data for the pilot sample

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10	Var 11	Var 12	Var 13	Var 14	Var 15	Var 16	Var 17	Var 18	Var 19	Var 20	Var 21	Var 22	Var 23	Var 24
2	M.D.	1	1	1	1	2	2	1	1	2	2	2	2	2	2	1	1	3	1	3	1	3	2
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	1	2	1	1	1	2	1	2	2	2	3	1	2	2	5	1	2	3
1	2	3	3	1	1	2	1	1	1	2	2	2	2	2	2	2	2	3	2	5	3	1	3
1	2	1	1	1	2	1	1	1	1	1	1	1	2	1	2	5	3	2	1	5	1	1	5
1	2	3	3	1	1	1	2	1	1	2	2	1	1	1	2	5	2	3	3	5	4	4	4
2	2	3	3	2	2	1	2	1	1	1	1	1	2	2	2	5	3	2	3	5	3	4	4
1	2	3	1	1	1	1	2	1	1	2	1	2	2	2	2	3	1	3	1	4	1	4	3
1	2	3	3	1	1	M.D.	M.D.	1	1	2	1	1	1	2	2	5	2	5	4	5	2	4	5
1	3	1	1	2	2	2	2	1	1	1	2	1	2	2	2	1	2	2	2	5	2	1	2
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	2	2	1	1	2	2	2	2	2	2	3	3	3	5	5	5	5	5
2	M.D.	2	2	2	2	1	2	1	1	2	1	1	1	2	2	3	4	2	1	3	3	2	1
2	1	2	2	2	2	2	2	1	1	1	2	1	2	1	2	4	3	3	1	5	3	4	5
1	M.D.	M.D.	M.D.	M.D.	1	2	2	1	1	1	2	1	1	2	2	5	3	3	3	3	3	3	2
1	2	2	2	1	M.D.	2	2	1	1	2	2	2	2	2	2	1	3	3	2	5	1	4	4
1	1	2	2	1	1	1	1	1	1	1	2	1	2	2	2	5	1	2	2	5	2	4	4
1	M.D.	2	2	2	2	2	2	1	1	2	2	1	2	2	2	3	2	4	3	2	3	2	3
2	3	2	2	2	2	2	2	1	1	1	1	2	2	2	2	3	2	1	1	5	4	5	5
1	3	2	2	1	1	M.D.	2	1	1	2	2	1	2	2	2	5	M.D.	1	1	4	1	1	1
2	NA	M.D.	2	2	1	1	2	1	1	1	1	1	1	1	2	5	4	3	2	4	3	3	4
1	2	3	3	1	1	2	2	1	1	1	2	1	2	1	2	3	3	3	3	5	4	4	4
1	3	3	3	2	2	2	2	1	1	1	1	1	2	2	2	4	1	3	1	5	1	3	5
2	M.D.	2	2	1	2	2	2	1	1	2	2	2	2	2	2	2	1	1	1	5	2	3	1
2	M.D.	2	2	1	1	1	2	1	1	2	2	1	2	2	2	5	1	2	1	3	1	1	5
1	3	2	2	1	2	2	2	1	1	2	2	2	2	2	2	3	2	2	1	4	3	2	2
M.D.	M.D.	M.D.	M.D.	2	1	1	2	1	1	2	1	2	2	2	2	1	1	4	1	4	2	3	4
1	2	2	2	1	2	2	2	1	1	2	2	2	2	2	2	M.D.	M.D.	M.D.	M.D.	5	M.D.	M.D.	M.D.
1	M.D.	2	2	1	2	2	2	1	1	2	2	2	2	2	2	1	1	1	1	3	1	1	1
2	2	3	3	2	2	2	2	1	1	2	2	2	2	2	2	3	2	3	2	3	2	3	2
1	4	3	3	1	1	1	2	1	1	2	2	2	2	1	2	5	1	1	1	5	1	5	5
2	1	1	1	1	2	1	2	1	1	2	1	1	2	2	2	5	3	2	2	3	3	2	2
2	1	2	3	2	2	2	2	1	1	1	2	2	2	1	2	4	1	2	1	4	3	3	4
1	2	2	1	1	2	2	2	1	1	2	2	1	2	2	2	5	2	3	4	4	1	3	4
1	3	2	2	1	1	1	2	1	1	2	1	1	2	2	2	1	1	5	3	4	3	3	2
1	M.D.	3	3	1	2	2	2	1	1	2	2	2	2	2	2	2	2	1	2	5	2	1	5
1	2	3	3	1	1	2	2	1	1	1	2	1	2	2	2	3	5	3	5	5	3	2	3
2	2	3	3	2	1	2	2	1	1	2	1	2	2	1	2	1	1	1	1	3	1	3	3
1	3	3	3	1	1	1	2	1	1	1	1	1	2	1	2	5	4	3	4	5	2	2	4
1	3	3	3	1	1	1	1	2	2	2	2	2	2	2	2	M.D.	M.D.	M.D.	M.D.	5	5	M.D.	M.D.
2	1	2	2	2	2	2	2	1	2	1	2	1	2	2	2	M.D.	M.D.	4	M.D.	5	4	M.D.	M.D.
2	2	2	2	1	1	2	2	1	1	2	1	1	1	2	2	1	5	4	5	4	5	5	4
2	M.D.	3	3	M.D.	M.D.	1	2	1	1	1	1	1	2	1	2	5	5	5	5	5	5	5	5
1	3	2	2	1	2	2	2	1	1	1	2	1	1	2	2	5	4	3	2	5	1	3	3
2	2	3	3	2	2	2	2	1	1	2	2	2	2	2	2	3	1	1	1	3	3	3	1

M.D.: Missing Data

Var1: Have you graduated from college of education?
Var2: Your level of graduation
Var3: No. of Years of Experience
Var4: No. of Years of Experience in secondary school
Var5: Do you utilise the instructional behavioural objectives designed by the MoE's recent project?
Var6: When writing your tests, do you do a specification table?
Var7: Have you ever attended a training course on tests' construction?
Var8: Have you ever attended a training course on test results' analysis?
Var9: Do you know how to use computers?
Var10: Microsoft Word®.
Var11: Microsoft Explorer®.
Var12: Microsoft Excel®.
Var13: Microsoft PowerPoint®.
Var14: Microsoft Access®.
Var15: Microsoft Outlook®.
Var16: SPSS®.
please describe the extent to which you are using a computer for each purpose:
Var17: As a word processor.
Var18: To manage my personal financial affairs.
Var19: For presenting (lessons, training courses … etc).
Var20: For scheduling tasks, appointments … etc.
Var21: Browsing and searching the Internet.
Var22: Doing calculations (other than personal).
Var23: Searching databases for personal purposes (such as telephone directories, names or addresses of people, organisations or products … etc)
Var24: Searching databases for non-personal purposes (such as electronic encyclopaedias, dictionaries and reference books … etc)

The pre and post-tests' raw data for the pilot sample

Var1	Var2	Var3	Var4	Var5	Var6
4	7	1	8		6
M.D.	11	M.D.	8	8	9
8	9	6	7	8	7
7		M.D.	9		3
9	10	7	10	5	10
4	6	6	8	7	7
8	8	M.D.	9	9	9
7	8	8	7	8	8
8	4	7	9	6	8
1	4	5	8	7	7
	11		8	7	6
	8	3	8	6	10
7	7	4	9	3	7
7	10	M.D.	7	7	9
6	9	7	8	7	8
6	5	5	7	10	6
M.D.		M.D.	7	5	10
7	6	M.D.	9	8	8
1	5	M.D.	7	7	9
5	5	4	8	7	7
6	5	7	6	3	10
4	7	M.D.	7	9	8
4	7	3	6	9	7
6		8	8		10
8	8	1	9	8	7
7	2	8	9	5	7
6	5	6	9	7	9
4	8	6	M.D.	M.D.	M.D.
9	7	9	10	8	10
4	6	2	8	5	8
5	4	M.D.	8	7	9
8	7	7	9	7	9
6	4	M.D.	8	9	8
2	5	1	8	5	8
8	9	7	M.D.	M.D.	M.D.
M.D.	9	4	M.D.	M.D.	M.D.
8	7	M.D.	M.D.	M.D.	M.D.
8		M.D.	4	M.D.	6
5	2	9	9	M.D.	8
7		M.D.	M.D.	M.D.	M.D.
7		10	8	8	10
1		1	7	7	7
M.D.	7	M.D.	7	8	7

M.D.: Missing Data

Pre-Test

Var1: Pre-test HCD knowledge
Var2: Pre-test HCD question construction skills
Var3: Pre-test IAT knowledge

Post-Test

Var4: Post-test HCD knowledge
Var5: Post -test HCD question construction skills
Var6: Post -test IAT knowledge

The self-report questionnaire's raw data for the pilot sample– Part 1

Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12
.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	2.00	.00	.00	.00	.00	.00	.00	.00	.00	1.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	2.00	2.00	.00	1.00	1.00	.00	1.00	.00	1.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	2.00	.00	1.00	.00	.00	.00	.00	.00	1.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
2.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	2.00	.00	.00	.00	.00	.00	.00	.00	1.00	.00
M.D.	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	2.00	M.D.	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	2.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
2.00	2.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	2.00	.00	1.00	1.00	1.00	.00	.00	1.00	1.00	.00
1.00	M.D.	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	2.00	.00	.00	.00	.00	.00	1.00	.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	2.00	1.00	.00	.00	1.00	.00	.00	.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	2.00	2.00	.00	.00	1.00	.00	.00	.00	.00	1.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	2.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	2.00	2.00	.00	1.00	1.00	1.00	.00	.00	.00	1.00	1.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	2.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
1.00	1.00	2.00	.00	.00	.00	.00	.00	.00	.00	.00	1.00
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
1.00	2.00	2.00	.00	.00	1.00	.00	.00	.00	.00	.00	.00
2.00	2.00	2.00	.00	.00	.00	.00	.00	.00	.00	1.00	.00
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
1.00	1.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

M.D.: Missing Data

Var1: Do you have a personal computer at home ?
Var2: If you have installed the CAIAT software on a PC, have you tried it out for discovering the quality of your tests' questions?

Var3: After attending the training course, have your use of HCD instructional behavioural objectives increased?
Var4: Not interested in this issue.
Var5: Do not have much time.
Var6: This adds to my work load.
Var7: The school does not appreciate such improvement in my ability.
Var8: I feel afraid that I might do some scientific errors if I tackled HCD, thus I tend to be limited to the lower cognitive level.
Var9: I think that teaching physics should not go further to this level of higher cognitive demand.
Var10: I did not understand how can I apply HCD concept implications in the real world.
Var11: Although I understand what I have learnt in this course, I think that I need some more time until I understand thoroughly and be able to apply it.
Var12: Other reasons

The self-report questionnaire's raw data for the pilot sample - Part 2

Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10
.0	.0	.0	.0	.0	.0	M.D.	M.D.	M.D.	M.D.
2.0	.0	.0	.0	.0	.0	.0	.0	1.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	1.0	1.0	.0	1.0	.0	1.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
M.D.	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	1.0	1.0	1.0	.0	.0	.0	1.0	1.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	.0	.0	.0	.0	1.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	.0	1.0	.0	.0	1.0	.0	1.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	1.0	1.0	.0	.0	.0	.0	1.0	1.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
2.0	.0	.0	.0	.0	.0	.0	1.0	.0	.0
2.0	.0	.0	.0	.0	.0	.0	.0	1.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0

M.D.: Missing Data

Var1: **After attending the training course, have your questions in the level of HCD, either during instruction, in worksheets, in home works or in your tests, increased?**

Var2: Not interested in this issue.

Var3: Do not have much time.
Var4: This adds to my work load.
Var5: The school does not appreciate such improvement in my ability.
Var6: I feel afraid that I might do some scientific errors if I tackled HCD, thus I tend to be limited to the lower cognitive level.
Var7: I think that teaching physics should not go further to this level of higher cognitive demand.
Var8: I did not understand how can I apply HCD concept implications in the real world.
Var9: Although I understand what I have learnt in this course, I think that I need some more time until I understand thoroughly and be able to apply it.
Var10: Other reasons

The self-report questionnaire's raw data for the pilot sample - Part 3

Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
2.0	.0	.0	.0	.0	.0	.0	.0	1.0	.0	1.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	1.0	1.0	.0	.0	1.0	.0	.0	1.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	1.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
M.D.	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	1.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	1.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	1.0
2.0	.0	1.0	.0	.0	.0	.0	.0	.0	.0	1.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	.0	1.0	1.0	.0	.0	.0	1.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	1.0	.0	.0	.0	1.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	.0	.0	.0	.0	.0	.0	1.0	.0	.0
2.0	.0	1.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	1.0	.0	.0	.0	.0	.0	.0	.0	.0
2.0	.0	1.0	.0	.0	.0	1.0	.0	1.0	.0	.0
2.0	1.0	1.0	.0	.0	.0	.0	.0	.0	.0	1.0
2.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
2.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	1.0
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
2.0	.0	1.0	.0	.0	.0	1.0	.0	.0	.0	.0
2.0	.0	.0	.0	.0	.0	.0	.0	1.0	.0	.0
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.	M.D.
1.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0

M.D.: Missing Data

Var1: **After you have attended the training course of HCD-CAIAT project, have you tried to use what you had learnt about the CAIAT software on a self-learn basis?**

Var2: Not interested in this issue.
Var3: Do not have much time.
Var4: I am confident that my due abilities do not need improvement
Var5: The school does not appreciate such improvement in my ability.
Var6: I did not understand what this project is all about.
Var7: This adds to my work load.
Var8: I did not understand how can I apply HCD concept implications in the real world.
Var9: Although I understand what I have learnt in this course, I think that I need some more time until I understand thoroughly and be able to apply it.
Var10: I don't see that this is a suitable method for improving the process of question construction
Var11: Other reasons

Appendix 12

The updated items of the data collection instruments as a result of the judges' opinions

Instrument – section	Item (Initial text)	Updated text of the item
Questionnaire/Pre-test	Questions 1-4 were not present	Questions 1-4 added
Pre-test: Section 1	3 – Higher cognitive demand (HCD) levels are: A- the top two levels of Bloom's Taxonomy B- the top three levels of Bloom's Taxonomy C- the top four levels of Bloom's Taxonomy D- the top five levels of Bloom's Taxonomy	3 – The highest level of cognitive demand of Bloom taxonomy is: A – Comprehension. B – Understanding. C – Mastery. D – Evaluation.
	5 – Giving an example is one aspect of fulfilling application level.	5 – Explaining a concept is one aspect of fulfilling application level.
Pre-test: Section 2	6 – If difficulty coefficient value is high then this means that the question is an easy question	6 – For a given question, if its difficulty coefficient value is 95% and its discrimination coefficient is 5% then this indicate that it is a good question.
	7 – The optimum value of discrimination index is 0.	7 – Item analysis technique aids us to judge quality of testing items upon pupils' responses to those items rather than the subjective opinion of an individual who will be looking at that test.

Appendix 13

SPSS output for calculations of reliability coefficient of the research questionnaire, Guttmann Split-half method – Pilot Sample

```
Method 2 (covariance matrix) will be used for this analysis ******

R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (S P L I T)
```

1. Q1 Do you have a personal computer at home?
2. Q2 Have you installed the CAIAT software on your own PC?
3. Q3 If you have installed the CAIAT software on a Pc, have you tried it out for discovering the quality of your tests' questions? (Regardless of you being have succeeded in this attempt(s) or not)
4. Q4 After attending the training course, have your use of HCD instructional behavioural objectives for instruction increased?
5. Q5 After attending the training course, have your questions in the level of HCD, either during instruction, in worksheets, in home works or in your tests, increased?
6. Q6 After you have attended the training course of HCD-CAIAT project, have you tried to use what you had learnt about the CAIAT software on a self-learn basis?

		Mean	Std Dev	Cases
1.	Q1	1.0857	.2840	35.0
2.	Q2	1.2571	.4434	35.0
3.	Q3	1.3429	.4816	35.0
4.	Q4	1.2571	.4434	35.0
5.	Q5	1.5143	.5071	35.0

N of Cases = 35.0

Statistics for	Mean	Variance	Std Dev	N of Variables
Part 1	3.6857	.6924	.8321	3
Part 2	2.7714	.5933	.7702	2
Scale	6.4571	2.1378	1.4621	5

Item Means	Mean	Minimum	Maximum	Range	Max/Min	Variance
Part 1	1.2286	1.0857	1.3429	.2571	1.2368	.0171
Part 2	1.3857	1.2571	1.5143	.2571	1.2045	.0331
Scale	1.2914	1.0857	1.5143	.4286	1.3947	.0242

Item Variances	Mean	Minimum	Maximum	Range	Max/Min	Variance
Part 1	.1697	.0807	.2319	.1513	2.8750	.0063
Part 2	.2269	.1966	.2571	.0605	1.3077	.0018
Scale	.1926	.0807	.2571	.1765	3.1875	.0046

Intraclass Correlation Coefficient

```
Two-Way Mixed Effect Model (Consistency Definition):
People Effect Random, Measure Effect Fixed
 Single Measure Intraclass Correlation =    .3050*
    95.00% C.I.:            Lower =    .1593          Upper =    .4855
 F =   3.1940   DF = (    34,  136.0)   Sig. = .0000  (Test Value = .0000 )
 Average Measure Intraclass Correlation =    .6869**
    95.00% C.I.:            Lower =    .4865          Upper =    .8251
 F =   3.1940   DF = (    34,  136.0)   Sig. = .0000  (Test Value = .0000 )
*: Notice that the same estimator is used whether the interaction effect
   is present or not.
**: This estimate is computed if the interaction effect is absent,
    otherwise ICC is not estimable.
```

Reliability Coefficients	5 items		
Correlation between forms =	.6647	Equal-length Spearman-Brown =	.7986
Guttman Split-half =	**.7972**	Unequal-length Spearman-Brown =	.8040
Alpha for part 1 =	.3968	Alpha for part 2 =	.4703
3 items in part 1		2 items in part 2	

Appendix 14

SPSS output for calculations of internal consistency coefficient of the questionnaire's items and reliability (Alpha) of the scale before correction - Pilot Sample

```
**** Method 2 (covariance matrix) will be used for this analysis ****
 R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)
```

1. Q1 Do you have a personal computer at home?
2. Q2 Have you installed the CAIAT software on your own PC?
3. Q3 If you have installed the CAIAT software on a Pc, have you tried it out for discovering the quality of your tests' questions? (Regardless of you being have succeeded in this attempt(s) or not)
4. Q4 After attending the training course, have your use of HCD instructional behavioural objectives for instruction increased?
5. Q5 After attending the training course, have your questions in the level of HCD, either during instruction, in worksheets, in home works or in your tests, increased?
6. Q6 After you have attended the training course of HCD-CAIAT project, have you tried to use what you had learnt about the CAIAT software on a self-learn basis?

		Mean	Std Dev	Cases
1.	Q1	1.0857	.2840	35.0
2.	Q2	1.2571	.4434	35.0
3.	Q3	1.3429	.4816	35.0
4.	Q4	1.2571	.4434	35.0
5.	Q5	1.5143	.5071	35.0

Correlation Matrix

	Q1	Q2	Q3	Q4	Q5
Q1	1.0000				
Q2	.2869	1.0000			
Q3	-.0061	.2636	1.0000		
Q4	.0534	.2521	.6768	1.0000	
Q5	.0934	.5718	.3407	.3102	1.0000

```
 R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)

        N of Cases =        35.0
```

Statistics for Scale	Mean	Variance	Std Dev	N of Variables
	6.4571	2.1378	1.4621	5

Item Means	Mean	Minimum	Maximum	Range	Max/Min	Variance
	1.2914	1.0857	1.5143	.4286	1.3947	.0242

Item Variances	Mean	Minimum	Maximum	Range	Max/Min	Variance
	.1926	.0807	.2571	.1765	3.1875	.0046

Inter-item Correlations	Mean	Minimum	Maximum	Range	Max/Min	Variance
	.2843	-.0061	.6768	.6830	-110.1682	.0438

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
Q1	5.3714	1.9462	.1400	.0968	**.7288**
Q2	5.2000	1.4000	.5157	.3894	.6034
Q3	5.1143	1.3395	.5081	.4818	.6056
Q4	5.2000	1.4000	.5157	.4682	.6034
Q5	4.9429	1.2908	.5120	.3724	.6042

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	14.5371	34	.4276		
Within People	21.6000	140	.1543		
Between Measures	3.3943	4	.8486	6.3390	.0001
Residual	18.2057	136	.1339		
Total	36.1371	174	.2077		
Grand Mean 1.2914					

Intraclass Correlation Coefficient

```
Two-Way Mixed Effect Model (Consistency Definition):
People Effect Random, Measure Effect Fixed
 Single Measure Intraclass Correlation =    .3050*
    95.00% C.I.:            Lower =    .1593          Upper =    .4855
 F =   3.1940   DF = (    34,  136.0)   Sig. = .0000  (Test Value = .0000 )
 Average Measure Intraclass Correlation =    .6869**
    95.00% C.I.:            Lower =    .4865          Upper =    .8251
 F =   3.1940   DF = (    34,  136.0)   Sig. = .0000  (Test Value = .0000 )
*: Notice that the same estimator is used whether the interaction effect
   is present or not.
**: This estimate is computed if the interaction effect is absent,
    otherwise ICC is not estimable.
```

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

Reliability Coefficients 5 items

Alpha = **.6869** Standardized item alpha = .6651

Appendix 15

SPSS output for calculations of internal consistency coefficient of the questionnaire's items and reliability (Alpha) of the scale after correction (by deletion of item No. 1) - Pilot Sample

```
**** Method 2 (covariance matrix) will be used for this analysis ****
 R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   (A L P H A)
```

2. Q2 Have you installed the CAIAT software on your own PC?

3. Q3 If you have installed the CAIAT software on a Pc, have you tried it out for discovering the quality of your tests' questions? (Regardless of you being have succeeded in this attempt(s) or not)

4. Q4 After attending the training course, have your use of HCD instructional behavioural objectives for instruction increased?

5. Q5 After attending the training course, have your questions in the level of HCD, either during instruction, in worksheets, in home works or in your tests, increased?

6. Q6 After you have attended the training course of HCD-CAIAT project, have you tried to use what you had learnt about the CAIAT software on a self-learn basis?

		Mean	Std Dev	Cases
1.	Q2	1.2500	.4392	36.0
2.	Q3	1.3333	.4781	36.0
3.	Q4	1.2500	.4392	36.0
4.	Q5	1.5000	.5071	36.0

Correlation Matrix

	Q2	Q3	Q4	Q5
Q2	1.0000			
Q3	.2722	1.0000		
Q4	.2593	.6804	1.0000	
Q5	.5774	.3536	.3208	1.0000

N of Cases = 36.0

Statistics for Scale	Mean	Variance	Std Dev	Variables
	5.3333	1.9429	1.3939	4

Item Means	Mean	Minimum	Maximum	Range	Max/Min	Variance
	1.3333	1.2500	1.5000	.2500	1.2000	.0139

Item Variances	Mean	Minimum	Maximum	Range	Max/Min	Variance
	.2179	.1929	.2571	.0643	1.3333	.0010

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
Q2	4.0833	1.2786	.4747	.3403	.7039
Q3	4.0000	1.1429	.5590	.4841	.6562
Q4	4.0833	1.2214	.5445	.4717	.6667
Q5	3.8333	1.1143	.5338	.3783	.6731

_

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	17.0000	35	.4857		
Within People	15.0000	108	.1389		
Between Measures	1.5000	3	.5000	3.8889	.0111
Residual	13.5000	105	.1286		
Total	32.0000	143	.2238		

Grand Mean 1.3333

Intraclass Correlation Coefficient

Two-Way Mixed Effect Model (Consistency Definition):
People Effect Random, Measure Effect Fixed
Single Measure Intraclass Correlation = .4098*
95.00% C.I.: Lower = .2407 Upper = .5913
F = 3.7778 DF = (35, 105.0) Sig. = .0000 (Test Value = .0000)
Average Measure Intraclass Correlation = .7353**
95.00% C.I.: Lower = .5591 Upper = .8526
F = 3.7778 DF = (35, 105.0) Sig. = .0000 (Test Value = .0000)
*: Notice that the same estimator is used whether the interaction effect is present or not.
**: This estimate is computed if the interaction effect is absent, otherwise ICC is not estimable.

Reliability Coefficients 4 items

Alpha = **.7353** Standardized item alpha = .7359

Appendix 16

Internal Consistency and Reliability of the scale.

Calculations results of internal consistency coefficient of the questionnaire's items and reliability (Alpha) of the scale.
Calculated by non-parametric correlation (Spearman – 2 tail) - Pilot Sample

Items mentioned by index number		2	3	4	5
1	Correlation Coefficient	**0.262**	**0.000**	**1.058**	**0.101**
	Sig. (2-tailed)	**0.123**	**1**	**0.737**	**0.560**
	N	**36**	**36**	**36**	**36**
2	Correlation Coefficient		**0.272**	**0.259**	**0.577 ****
	Sig. (2-tailed)		**0.108**	**0.127**	**0.000**
	N		**36**	**36**	**36**
3	Correlation Coefficient			**0.684 ****	**0.365 ***
	Sig. (2-tailed)			**0.000**	**0.026**
	N			**37**	**37**
4	Correlation Coefficient				**0.330 ****
	Sig. (2-tailed)				**0.046**
	N				**37**

Item Index	I t e m T e x t
1	**Do you have a personal computer at home?**
2	**If you have installed the CAIAT software on a Pc, have you tried it out for discovering the quality of your tests' questions? (Regardless of you being have succeeded in this attempt(s) or not?**
3	**After attending the training course, have your use of HCD instructional behavioural objectives increased?**
4	**After attending the training course, have your questions in the level of HCD, either during instruction, in worksheets, in home works or in your tests, increased?**
5	**After you have attended the training course of HCD-CAIAT project, have you tried to use what you had learnt about the CAIAT software on a self-learn basis?**

**** Correlation is significant at 0.05 level (2-tailed)***

Appendix 17

Tables of detailed findings for tests of statistical significance – Main Sample

Effectiveness Dimension

Appendix Table 17.1: Independent samples T-test for HCD concept's

	Targeting differences between individuals upon their:		**Average**	**no of Cases**	**Standard Deviation**	**Degrees of Freedom**	**T value**	**Significance**
Pre-test	*educational qualification*	**Yes**	4.6585	246	3.00927	**326**	**-2.107**	**0.036 ***
		No	3.8659	82	2.76112			
	training on test construction	**Yes**	4.7474	194	2.87442	**337**	**-2.089**	**0.037 ***
		No	4.0759	145	2.99788			
	training on IAT	**Yes**	5.6286	35	2.57917	**332**	**-2.490**	**0.013 ***
		No	4.3344	299	2.94386			
	key stage (intermediate/secondary)	**Inter.**	4.1018	167	2.90955	**344**	**-2.098**	**0.037 ***
		Sec.	4.7654	179	2.96813			
Post Test	*educational qualification*	**Yes**	7.9336	226	1.23289	**296**	**-0.753**	**0.452**
		No	7.8056	72	1.32834			
	training on test construction	**Yes**	7.8212	179	1.21849	**305**	**1.468**	**0.143**
		No	8.0313	128	1.26078			
	training on IAT	**Yes**	7.6970	33	1.61022	**301**	**0.830**	**0.412**
		No	7.9370	270	1.17586			
	key stage (intermediate/secondary)	**Inter.**	7.8523	149	1.22682	**312**	**-0.880**	**0.380**
		Secondary	7.9758	165	1.25403			

Appendix Table 17.2: One-way ANOVA test (F value) for HCD concepts

	Targeting the differences between individuals upon their:	Source of Variance	Sum of Squares	Degrees of Freedom	Average of Squares	F Value	Significance
Pre-test	*level of graduation*	Among groups	60.736	3	20.245	2.361	0.071
		Within groups	2855.178	333	8.574		
		Sum	2915.923	336			
	number of years of experience in teaching	Among groups	0.283	2	0.142	0.016	0.984
		Within groups	2944.253	336	8.763		
		Sum	2944.537	338			
	specialisation subject (physics – chemistry – biology).	Among groups	76.072	3	25.357	2.963	0.032 *
		Within groups	2935.691	343	8.559		
		Sum	3011.764	346			
Post-test	*level of graduation*	Among groups	5.510	3	1.837	1.180	0.318
		Within groups	468.601	301	1.557		
		Sum	474.111	304			
	number of years of experience in teaching	Among groups	9.003	2	4.502	2.939	0.054
		Within groups	467.127	305	1.532		
		Sum	476.130	307			
	specialisation subject (physics – chemistry – biology).	Among groups	6.975	3	2.325	1.519	0.210
		Within groups	476.041	311	1.531		
		Sum	483.016	314			

Appendix Table 17.3: Independent samples T-test for writing HCD questions

	Targeting differences between individuals upon their:		Average	no of Cases	Standard Deviation	Degrees of Freedom	T value	Significance
Pre-test	*educational qualification*	Yes	4.5949	237	1.95567	314	-0.198	0.844
		No	4.5443	79	2.02420			
	Training on test construction	Yes	4.5699	186	1.92226	324	0.072	0.942
		No	4.5857	140	1.99184			
	training on IAT	Yes	4.8788	33	1.99621	320	-0.951	0.343
		No	4.5329	289	1.97856			
	key stage (intermediate/secondary)	Inter.	7.8713	167	16.92515	344	-0.083	0.934
		Sec.	8.0196	179	16.28817			
Post Test	*educational qualification*	Yes	6.4031	227	1.81667	297	-0.672	0.502
		No	6.2361	72	1.90065			
	Training on test construction	Yes	6.2095	179	1.94863	306	1.709	0.089
		No	6.5659	129	1.58654			
	training on IAT	Yes	6.1061	33	1.90705	302	0.865	0.388
		No	6.3967	271	1.81108			
	key stage (intermediate/secondary)	Inter.	5.9733	150	1.99982	313	-3.852	0.001 *
		Sec.	6.7424	165	1.53079			

Appendix Table 17.4: One-way ANOVA test (F value) for writing HCD questions

	Targeting the differences between individuals upon their:	Source of Variance	Sum of Squares	Degrees of Freedom	Average of Squares	F Value	Significance
Pre-test	*level of graduation*	Among groups	29.807	3	9.936	2.580	0.054
		Within groups	1228.585	319	3.851		
		Sum	1258.392	322			
	number of years of experience in teaching	Among groups	7.420	2	3.710	0.959	0.384
		Within groups	1254.772	322	3.869		
		Sum	1253.192	324			
	specialisation subject (physics – chemistry – biology).	Among groups	232.828	3	77.609	24.142	0.000 *
		Within groups	1057.651	329	3.215		
		Sum	1290.479	332			
Post-test	*level of graduation*	Among groups	10.539	3	3.513	1.122	0.340
		Within groups	954.259	302	3.130		
		Sum	955.798	305			
	number of years of experience in teaching	Among groups	10.789	2	5.395	1.656	0.193
		Within groups	996.945	306	3.258		
		Sum	1007.735	308			
	specialisation subject (physics – chemistry – biology).	Among groups	21.693	3	7.231	2.245	0.083
		Within groups	1005.120	312	3.222		
		Sum	1026.812	315			

Appendix Table 17.5: Independent samples T-test for IAT skills

	Targeting differences between individuals upon their:		Average	no of Cases	Standard Deviation	Degrees of Freedom	T value	Significance
Pre-test	*educational qualification*	Yes	5.0772	246	2.78299	326	-1.190	0.235
		No	4.6585	82	2.68613			
	Training on test construction	Yes	5.3505	194	2.58550	337	-2.687	0.008 *
		No	4.5379	145	2.87475			
	training on IAT	Yes	5.6286	35	2.01590	332	-1.816	0.075
		No	4.9431	299	2.81059			
	key stage (intermediate/secondary)	Inter.	4.9102	167	2.66577	344	-0.457	0.648
		Sec.	5.0447	179	2.80213			
Post Test	*educational qualification*	Yes	8.1637	226	1.20912	296	-1.207	0.228
		No	7.9722	72	1.04776			
	Training on test construction	Yes	8.1508	179	1.11911	305	-0.303	0.762
		No	8.1094	128	1.26885			
	training on IAT	Yes	8.1818	33	1.04447	301	-0.342	0.733
		No	8.1074	270	1.19461			
	key stage (intermediate/secondary)	Inter.	7.8591	149	1.34576	312	-3.975	0.001 *
		Sec.	8.3758	165	0.93933			

Appendix Table 17.6: One-way ANOVA test (F value) for IAT skills

	Targeting the differences between individuals upon their:	**Source of Variance**	**Sum of Squares**	**Degrees of Freedom**	**Average of Squares**	**F Value**	**Significance**
Pre-test	*level of graduation*	**Among groups**	**34.611**	**3**	**11.537**	**1.582**	**0.194**
		Within groups	**2428.807**	**333**	**7.294**		
		Sum	**2463.418**	**336**			
	number of years of experience in teaching	**Among groups**	**2.877**	**2**	**1.438**	**0.193**	**0.825**
		Within groups	**2503.979**	**336**	**7.452**		
		Sum	**2506.855**	**338**			
	specialisation subject (physics – chemistry – biology).	**Among groups**	**2.480**	**3**	**0.827**	**0.110**	**0.954**
		Within groups	**2585.474**	**343**	**7.538**		
		Sum	**2587.954**	**346**			
Post Test	*level of graduation*	**Among groups**	**10.262**	**3**	**3.421**	**2.489**	**0.060**
		Within groups	**413.587**	**301**	**1.374**		
		Sum	**423.849**	**304**			
	number of years of experience in teaching	**Among groups**	**4.587**	**2**	**2.293**	**1.636**	**0.196**
		Within groups	**427.475**	**305**	**1.402**		
		Sum	**432.062**	**307**			
	specialisation subject (physics – chemistry – biology).	**Among groups**	**5.995**	**3**	**1.998**	**1.449**	**0.229**
		Within groups	**428.926**	**311**	**1.379**		
		Sum	**434.921**	**314**			

Adoption Dimension

Appendix Table 17.7: Independent samples T-test for teachers' averages of trying out the CAIAT software on a self-learning basis with respect to the differences of their background on different factors

Targeting differences between individuals upon their:		Number of Cases	Average	T value	Significance
educational qualification	Yes	156	1.853	-0.224	0.823
	No	37	1.838		
Training on test construction	Yes	102	1.833	0.193	0.847
	No	77	1.844		
Training on IAT	Yes	20	1.950	-2.017	0.051
	No	155	1.832		
Key stage (intermediate/secondary)	Intermediate	91	1.813	-1.170	0.244
	Secondary	111	1.873		
Do you know how to use computers?	Yes	141	1.825	-0.287	0.774
	No	40	1.844		
Do you have a PC at home?	Yes	179	1.849	-0.378	0.706
	No	22	1.818		
Can you use Excel software?	Yes	69	1.840	0.162	0.872
	No	113	1.850		
Can you use Access software?	Yes	15	1.866	-0.229	0.819
	No	167	1.844		

Appendix Table 17.8: ANOVA test (F value) for teachers' averages of trying out the CAIAT software on a self-learning basis with respect to the differences of their background on different factors

Targeting the differences between individuals upon their:	Source of Variance	Sum of Squares	Degrees of Freedom	Average of Squares	F Value	Significance
Level of graduation	Among groups	0.371	3	0.124	0.928	0.428
	Within groups	25.750	193	0.133		
	Sum	26.122	196			
number of years of experience in teaching	Among groups	0.381	2	0.191	1.476	0.231
	Within groups	25.050	194	0.129		
	Sum	25.431	196			
specialisation subject (physics – chemistry– biology)	Among groups	0.307	3	0.102	0.782	0.505
	Within groups	25.935	198	0.131		
	Sum	26.243	201			

Appendix Table 17.9: Independent samples T-test for teachers' use of HCD instructional objectives with respect to the differences of their background on different factors

Targeting differences between individuals upon their:		Number of Cases	Average	T value	Significance
educational qualification	Yes	168	1.946	0.730	0.466
	No	39	1.974		
Training on test construction	Yes	114	1.947	0.061	0.951
	No	79	1.949		
Training on IAT	Yes	23	1.957	-0.209	0.835
	No	167	1.946		
Key stage (intermediate/secondary)	Intermediate	96	1.927	-1.274	0.204
	Secondary	121	1.967		
Do you know how to use computers?	Yes	150	1.96	-1.067	0.290
	No	45	1.911		
Do you have a PC at home?	Yes	192	1.953	-0.763	0.446
	No	24	1.917		
Can you use Excel software?	Yes	72	1.944	0.501	0.617
	No	125	1.96		
Can you use Access software?	Yes	15	2.000	-0.879	0.381
	No	182	1.951		

Appendix Table 17.10: ANOVA test (F value) for teachers' use of HCD instructional objectives with respect to the differences of their background on different factors

Targeting the differences between individuals upon their:	Source of Variance	Sum of Squares	Degrees of Freedom	Average of Squares	F Value	Significance
Level of graduation	Among groups	0.172	3	0.057	1.160	0.326
	Within groups	10.258	208	0.049		
	Sum	10.429	211			
number of years of experience in teaching	Among groups	0.052	2	0.026	0.574	0.564
	Within groups	9.474	208	0.046		
	Sum	9.526	210			
specialisation subject (physics –chemistry– biology)	Among groups	0.093	3	0.031	0.641	0.590
	Within groups	10.349	213	0.049		
	Sum	10.442	216			

Appendix Table 17.11: Independent samples T-test for teachers' questions of HCD level during instruction with respect to the differences of their background on different factors

Targeting differences between individuals upon their:		Number of Cases	Average	T value	Significance
educational qualification	Yes	166	1.964	2.487	0.014 *
	No	39	2.000		
Training on test construction	Yes	114	1.974	000	1.000
	No	76	1.974		
Training on IAT	Yes	23	2.000	-0.846	0.399
	No	164	1.970		
Key stage (intermediate/secondary)	Intermediate	95	1.957	-0.688	0.492
	Secondary	119	1.974		
Do you know how to use computers?	Yes	149	1.973	0.130	0.897
	No	43	1.977		
Do you have a PC at home?	Yes	190	1.968	-0.301	0.764
	No	23	1.957		
Can you use Excel software?	Yes	72	1.972	0.135	0.893
	No	122	1.975		
Can you use Access software?	Yes	15	2.000	-0.653	0.514
	No	179	1.972		

Appendix Table 17.12: ANOVA test (F value) for teachers' questions of HCD level during instruction with respect to the differences of their background on different factors

Targeting the differences between individuals upon their:	Source of Variance	Sum of Squares	Degrees of Freedom	Average of Squares	F Value	Significance
Level of graduation	Among groups	0.246	3	0.082	2.580	0.055
	Within groups	6.519	205	0.032		
	Sum	6.766	208			
number of years of experience in teaching	Among groups	0.019	2	0.009	0.282	0.755
	Within groups	6.746	205	0.033		
	Sum	6.764	207			
specialisation subject (physics –chemistry– biology)	Among groups	0.039	3	0.013	0.408	0.747
	Within groups	6.732	210	0.032		
	Sum	6.771	213			

Appendix 18

Qualitative Findings of the Pilot Sample

Appendix Table 18.1: Findings of observation during the workshops for the pilot sample

Area	#	Observational Guides	First Workshop		Second Workshop	
			Yes	No	Yes	No
Running CAIAT Software	1	Do they have the ability to run the software?	96.3%	3.7%	100%	-
	2	Do they know the meanings of its menus' components?	77.8%	22.2%	100%	-
	3	Do they know how to enter the initial specification of the test?	81.5%	14.8%	95.7%	4.3%
	4	Do they know how to enter marks to the software?	96.3%	3.7%	95.7%	4.3
	5	Can they review entered information easily?	88.9%	11.1%	81.8%	18.2%
	6	Can they edit any data entry errors easily?	96.3%	3.7%	65.2%	34.8%
	7	Do they read through the printed reports easily?	90%	10%	90.9%	9.1%
Skills of IAT	8	Do they possess enough ability to single out the suspected weak test's items upon their analyses results?	81%	19%	93.8%	6.2%
	9	Do they possess enough ability to single out the good test's items upon their analyses results?	85.7%	14.3%	90%	10%
	10	Do they have enough ability to find out what is the probable cause of the suspected weakness for an item?	65%	35%	81.5%	18.5%
	11	Do they have the ability to refine those weak items upon criteria of optimum test questions?	61.9%	38.1%	86.2%	13.8%
	12	Do they find any difficulty in comparing two coefficients to each other to gain one judgement? (For example difficulty coefficient versus discrimination index).	30%	70%	22%	78%

Appendix Table 18.2: Findings of questionnaire of the pilot sample's teachers' open-ended evaluation

Response Category	Response	F	P
Benefits?	1. Discovering weaknesses in testing items leads to an increase in my ability of writing good items in future.	**18**	**41%**
	2. It helps to ask questions in higher cognitive levels.	**8**	**18%**
	3. This helps in discovering questions' ability to give a true picture about pupils' level.	**4**	**9%**
	4. This aids in decreasing time and effort for item analysis process.	**3**	**7%**
	5. Knowing teachers from different schools and exchange experiences with them.	**2**	**5%**
	6. To continue self-development.	**1**	**2%**
Obstacles and Difficulties?	7. Data entry is time and effort consuming.	**10**	**23%**
	8. Inadequacy of the date of the training course.	**7**	**16%**
	9. After I finished from entering data and perform analyses, I was not able to add a new pupil to the class.	**3**	**7%**
	10. Difficulties in installing the software.	**3**	**7%**
	11. Pupils dislike questions that elicit thinking.	**2**	**5%**
	12. I had difficulties with printing.	**1**	**2%**
	13. When I change my mind about a question to be changed from subjective to objective or vise versa, I was not able to do so.	**1**	**2%**
	14. There is no way to transfer initial data (subjects and pupils names) to another file.	**1**	**2%**
	15. Inadequacy of the place of the training course.	**1**	**2%**
	16. Generally I am not qualified enough in using computers.	**1**	**2%**
	17. Principals do not appreciate such as this software.	**1**	**2%**
Suggestions?	18. HCD level theme needs to be given longer time.	**6**	**14%**
	19. I advise that the software be distributed to all teachers for all subjects.	**1**	**2%**
	20. This software is better to be linked with Ma'aref software.	**1**	**2%**
	21. Statistical analyses such as average and standard deviation to be added to the software outcomes.	**1**	**2%**
Feelings?	22. Expectation of the software's success in improving testing items construction and increasing levels of pupils.	**6**	**14%**
	23. Positive feelings during the training course.	**5**	**11%**
	24. Expectations of the project's failure when distributed further because it adds to the teacher workload.	**2**	**5%**
	25. Positive feelings due to the development in my performance.	**1**	**2%**
	26. Negative feelings (frustration) when I got analyses reports that raise a challenge to improve my testing items.	**1**	**2%**
	27. Enthusiasm to apply the CAIAT software for judging the quality of testing items.	**1**	**2%**
	28. Using the software support a sort of validity for testing items.	**1**	**2%**
Reflection on pupils?	29. Increasing pupils' role much more as a result of using thinking questions.	**4**	**9%**
	30. Increasing pupils' participation as a result of asking questions for thinking.	**4**	**9%**
	31. Increasing pupils' abilities to analyse.	**1**	**2%**
	32. This is convincing that the teacher could be behind failure of pupils as a result of lack of testing items quality.	**1**	**2%**

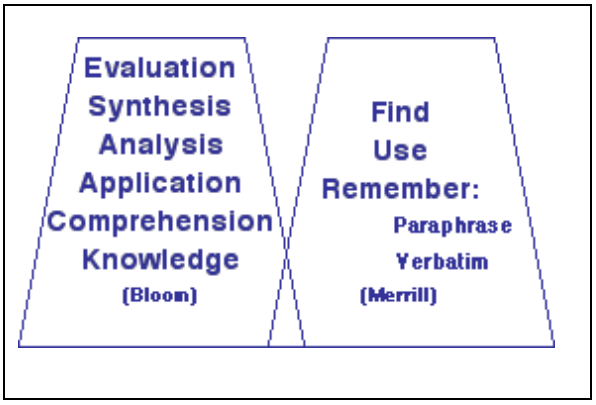
Endnotes

[1] A number of paragraphs and sections, especially the section of "Context" are adapted from my MA dissertation proposal report, an unpublished paper introduced to the University of Sussex Institute of Education during the summer semester 2001 under the supervision of John Pryor.

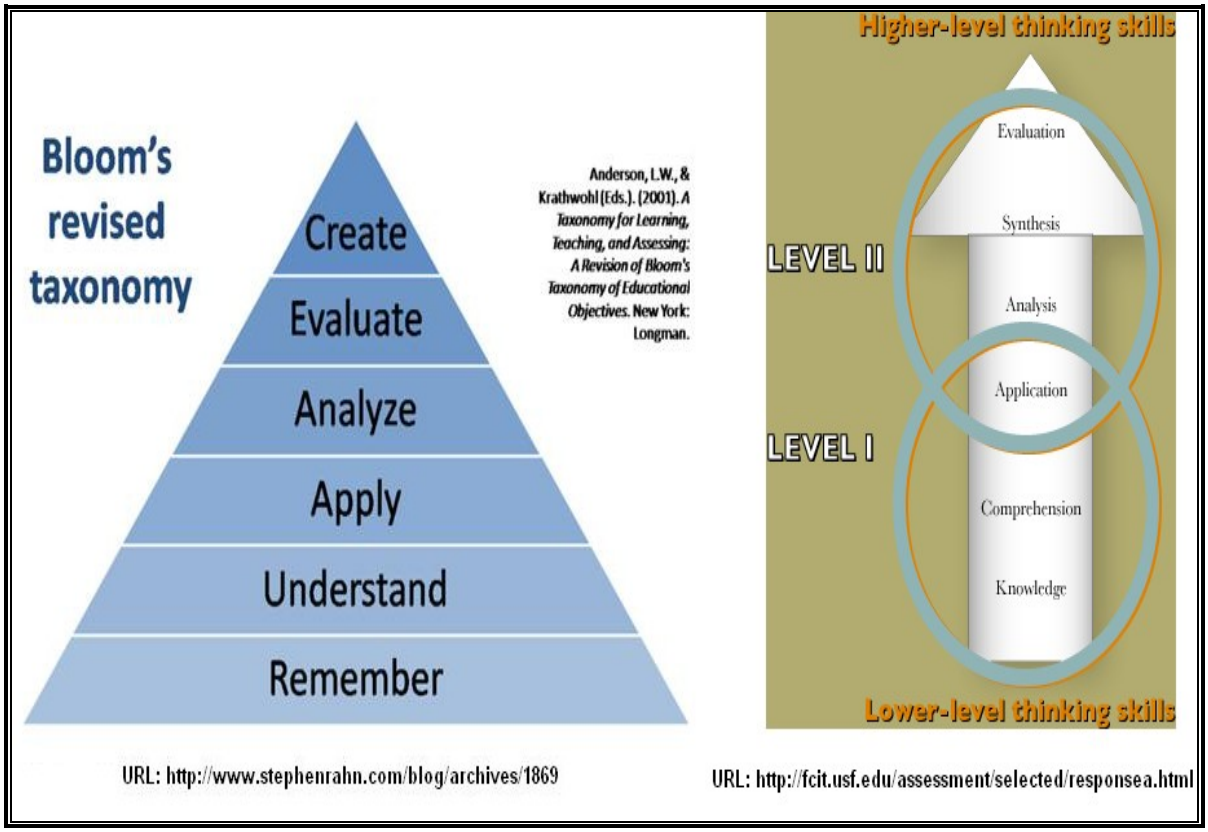
[2] For example, during the 1970s and '80s, most Saudi PG students sent to the West by the government studied in the USA and hence brought back the widespread trend of using multiple-choice test questions which was emerging in the USA at that time. My first-hand observation as a teacher and a supervisor working for the MoE since 1985 is that the effect of these university tutors on their teacher students resulted in new teachers coming to school, obviously focusing on using multiple-choice questions intensively. Although the Saudi education system has its own identity in many educational areas such as religion or language, it has made use of modern Western trends, ideas or methodologies to excel in other educational aspects that help to provide an optimal educational service involving such factors as curriculum design, instructional methods and assessment.

[3] There are hierarchies other than Bloom's: one example is Merrill's 3+ which is shown on the diagram below in comparison with Bloom's taxonomy. However, it is noted here that the two main categories of Bloom's taxonomy (LCD and HCD) could be represented by 'Remember' for the former and 'Use and Find' for the latter.

Comparison of Bloom and Merrill (Wolfe, Overbaugh and Bol, 2005)



Moreover, many more versions of the taxonomy have been suggested such as those shown below. However, in this research the emphasis is on the two major categories HCD and LCD, so it does not matter whether some of the levels were reorganised or substituted by other levels because of the general aspect of having HCD levels and LCD levels still present in all of these forms. The treatment of HCD levels in this research considers their nature being of higher thinking skills regardless of what these levels are.



[4] This theory (CTT) with the modern test theory (IRT) will be explained further in Chapter 4.

[5] For example, selecting the upper group and the lower group of pupils depends on their total scores (for all subjects), while their scores for the calculated subject may not reflect the same category, which for some is a source of confusion.

[6] Literal comprehension: the teacher provides the pupils with a direct explanation about the subject and hence expects that they will understand the related ideas. Inferential comprehension: the teacher provides the pupils with initial information about the subject and expects that they will think about this information before coming up with new ideas.

[7] Al Kharj is a KSA province.

[8] IRT is explained in Chapter 4.

[9] Although valid testing is mostly reliable, but not vice versa, reliability is a prerequisite to achieving validity and hence is one of its major potentials (Stanley and Hopkins, 1978, p.131).

[10] Has recently been renamed the General Directorate for Educational Evaluation and Quality.

[11] They distributed the first version of this item bank CD in 2001 to experimental schools, but sadly, the students at these schools were able to obtain the CD and hence memorised the answers as preparation for the test. This backward step eliminated achieving a paradigm shift to HCD level. I believe that if the item bank size was large enough to eliminate the value of memorising its test items then this leak of item bank CD would not have caused any significant contamination to the intervention's goals. However, the team is now commencing new trends and plans for another production. According to my meeting with Shukri, an MA student at Sussex who was part of a team for a similar project in Malaysia, they were constructing their item bank by selecting questions from actual teacher-made tests that had been examined in terms of validity and item analysis parameters. I believe this strategy is better, because it provides the educational authority with a huge number of testing items and it is thus less likely that the KSA experience would take place.

[12] There are a number of statistical models or calculating methods used for producing IAT parameters.

[13] The "self-instruction" program used in Tashkandi's study is similar in concept to the present study's method of teachers' self-learning.

[14] I have to indicate that two Vice-Ministers for girls' education have been in post since that Vice-Minister, which steers the direction of interest somehow to other issues according to each Vice-Minister's agenda for educational development. However, the interest in improving testing practice still exists and long-term efforts such as the present one are acknowledged.

[15] There are some Arabic IAT software packages developed by companies working in psychometric area but these are dedicated to be used for large-scale assessment with marking scanners thus cannot to be utilised by teachers.

[16] During the write-up of this research, and as from 2007, the national test as a central test was cancelled and the examination policy changed the assessment of the third year of high school to TA only.

[17] *Conservation* includes the idea that a quantity remains the same despite changes in appearance (two glasses of different shape and similar amount of water), *classification* is being able to classify objects logically and *seriation* is putting things in an adequate order.

[18] This term is written here as it appeared on the original. The language proof-reader recommended the dictionary form "smorgasbord."

[19] I am aware that Wikipedia is not a reliable source but the good presenting shape of these two diagrams as provided by Wikipedia with the fact that I am sure they are common in literature encouraged me to use them.

[20] It should be noted that, in KSA, school boards of the kind found in the West do not exist.

[21] Senior teachers are not part of the Saudi school staffing structure.

[22] Following the concept of the American Magnet School (1960s) the *Attracting School* is a concept also very similar to *Leading School* but focuses on hands-on knowledge and gives attention to making the school a place that is loved by the pupils and including three dimensions: school environment, school programme and instruction environment. The school environment is supported by providing the school with all the needed improvements/refurbishments for the building and facilities. The school programme includes changing the lesson time from 45 to 35 minutes, making three breaks throughout the school day, adding a daily 45-minute lesson dedicated to extra-curricular activities, focusing on field visits, and an open day every 8 weeks. The instruction environment includes providing effective school leadership, establishing an incentives principle, extending the learning environment to outside the classroom, more fun and enjoyment, applied learning instead of instruction, attention to extra-curricular activities, and utilising ICT in instruction. The evaluative study revealed that these schools' pupils' academic level was similar to other regular schools' levels. This failure rendered to shortcomings in providing the schools' needed improvements in the physical environment and educational supplies (Al-Awwad et al., 2010b). The point here is that by looking at opinions of pupils, parents, teachers, and managerial staff, none of these obstacles subscribed to people's negative attitudes to change; the problem was external as coming from the MoE's poor support.

[23] Rowntree (1987) introduced a third type which is *Ipsative*, or self-referenced assessment, in which each pupil compares the present performance with his or her own previous performance. This type is appropriate for a pupil's self-assessment (Freeman and Lewis, 1998: 21).

[24] This is according to what I have concluded above in the section of norm referencing and criterion referencing debates.

[25] Findley's D: The naming is due to an article by Warren G. Findley , "A Rationale for Evaluation of Item Discrimination Statistics" published in 1956 in the Journal of Educational and Psychological Measurement, volume 16, pp. 175-180 (Teaching & Learning Resource Centre, 2002).

[26] I used Visual Basic© 6 programming language to write the program code, Setup Factory© to compile the programming code and to create the executable file, and MS ACCESS© to create the database that holds the processed data.

[27] The furthest distance of these rural locations is almost 300 km (186 miles) and the nearest is about 50 km (31 miles). These locations are in deserts where there are no trains or public transportation to aid them to easily reaching there. Actually they use their own cars to travel daily and would have received no compensation for this at all.

[28] Intermediate schooling includes years 7 through 9 for ages 13 to 15. As for high schooling, it includes years 10 through 12 for ages 16 to 18.

[29] Women in KSA have no permission to drive cars. Also, no public transportation is available in KSA; therefore, female teachers depend mainly on their husbands, adult sons, private drivers, taxis, or fathers for those not married to drive them to school or any other place. During school time, the school bus drives the female teacher to the training venue or any other place related to her work.

[30] The figures here are a synopsis of what is illustrated in detail in the Tables of Appendix 17: Appendix Table 17.1 to Appendix Table 17.6.

[31] The T value and other numerical related values could be reviewed from Appendix Table 17.1. Also, in discussions of hypotheses that will follow I will shorten the format of the presentation into a form that will not include numbers, since they are shown in the designated table(s).

[32] This lesson observation is carried out by the teacher's educational supervisor who usually visits the teacher no less than twice a semester for performance evaluation and help her improve. Thus this project observation is widely applied because took place as part of their on-going work.

[33] It should be noted that the teachers were asked to write the entire test questions of HCD level as an experimental test, which is considered an exercise for their abilities so far.

[34] I mean that when the teacher utilises IAT, she makes use of her understanding of the background of the HCD concepts as part of the theoretical framework of the semi-action research she will perform. At the same time, this IAT provides her with lessons enriching her perception about her understanding of HCD skills. Many misunderstandings about HCD questioning corrected by the teachers' continuous minds-on exercise of IAT.

[35] I have to point out that a later observation that is made during the lessons is similar in function but different in purpose. The lesson observation is about HCD and measure the teachers' practices as a result of the whole project, while the workshop observation is about CAIAT/IAT and measures the immediate outcome of the training.

[36] To consolidate this finding, opinions from the pilot sample (male teachers) confirmed that the training was effectively encouraging them to participate actively. For example, one male teacher mentioned that:

> This training course differs from many training courses that I have attended before. It raises new trends and provides us with new ideas. The techniques that the course includes enable us to see the way in which we are going to apply it.

[37] Being qualitative data, the corresponding pilot sample findings are provided in Table 18.2 of Appendix 18 to compare the sort of problems raised or comments made by the two groups whenever necessary.

[38] Statistically, this is because the average of those not qualified is greater than that of qualified according to Appendix Table 17.11.

[39] I need to clarify why I listed the last two variables of the T-test table; namely: "Can you use Excel©?" and "Can you use Access©?" whereas there are other software packages that could be included similarly. The reason is that I considered these two as representatives of the mainstream software packages because Excel© is similar to the CAIAT as a calculations' software, and Access© is the database engine that the CAIAT software uses for storing its data..

[40] The figures here are a synopsis of what is illustrated in detail in the Tables of Appendix 17: Appendix Table 17.7 to Appendix Table 17.12.

[41] This issue appeared more clearly in the pilot sample's data of male teachers from Table 18.1 of Appendix 18 where the entering data percentage was 95.7% while the reviewing previous data and editing errors of data entry percentages were 81.8% and 65.2% respectively. This confirms the level of difficulty mentioned.

[42] This is embodied not only by item (b) "Do not have much time" but also by item (i) "Other Reasons: Because I have many subjects to teach", which was in the first order subscribes to the same notion.

[43] The technical reason is that there were some contradictions between some files of the CAIAT software and some other software package files available in some PCs. Because this contradiction depended on the other software packages installed previously on the machine, it was not under control thus needed technical support most of the time.

[44] The different problems that the software had at the pilot stage did not affect its main function since it provided the user with a facility to enter data, process it and then produce the required analyses. A second release for the main sample was much better. Nevertheless, problems remain even with the third release, but on a more specific levels, as items 6-19 and 27 of Table 6.18 reveal with some complaints having exaggerations (items 16-18). Obviously, this phenomenon seems to be normal in the software development field.

[45] When calculated, the whole range of responses adds up to 200.1% (there are repeated responses because each respondent could choose more than one reason). 32% of this ratio, i.e. of 200.1%, represents the percentage of the sum of all intrinsic reasons, while 68% of this ratio represents the sum of all the external reasons. These percentages could also be calculated similarly for Tables 7.2 and 7.3 and also show a similar low value for intrinsic reasons compared to that for external reasons.

[46] See section "Change in KSA" in Chapter 3.

[47] A similar aspect was seen with the pilot sample. Three male teachers attended a special evening session in basics of using computers though they were not paid for attendance. Some teachers, instead of calling the technical assistant to their school, brought their computers to his office.

[48] This means that Lertap© needs the presence of Excel© software on the PC in order to run.

[49] I used Visual Basic© 6 programming language to write the program code, Setup Factory© to compile the programming code and to create the executable file, and MS ACCESS© to create the database that holds the processed data.

[50] It should be noted that the high value of the difficulty coefficient indicates an easy question. This scheme is the most commonly used in related literature thus was chosen in this project.